USE OF THE VECTOR SPACE MODEL IN ENVIRONMENTAL SCANNING VIA THE WORLD WIDE WEB

By

CHERYL AASHEIM

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2002

ACKNOWLEDGMENTS

I would like to express my gratitude to the members of my committee, Drs. Gary J. Koehler, S. Selcuk Erenguc, Haldun Aytug and Jason Karceski. I greatly appreciate the insights, support and guidance provided by my committee chair, Dr. Koehler.

I owe the greatest debt to my parents for their continuous love, support and guidance. Their belief in me led to the belief I have in myself. I am grateful to my husband for his patience as well as his advice and constant support of anything that I want to pursue in life.

I would like to extend my gratitude to some of my fellow doctoral students Yi (Norman) Sun, Jonathan Smith, Nihat Kasap, Qian Tang, Mark Cecchini and Ling He for providing me with additional computing resources and expressing their support for my research.

TABLE OF CONTENTS

LIST OF TABLES

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

USE OF THE VECTOR SPACE MODEL IN ENVIRONMENTAL SCANNING VIA THE WORLD WIDE WEB

By

Cheryl Aasheim

August 2002

Chair: Dr. Gary J. Koehler
Major Department: Decision and Information Sciences

Environmental scanning is the process an organization uses to collect, analyze and use information. With the availability of vast quantities of information on the Internet, an organization has a great need for an automated methodology to scan and use this information. Additionally, the information available via the Internet is mostly text-based. Hence, the automated scanning methodology developed in this research uses the well-founded vector space model (VSM) to represent the documents available via the Internet and linear discriminant analysis to classify the documents. Chapter 1 of this dissertation provides an introduction to the environmental scanning problem in light of the current problem of gathering and using the vast quantity of information available via the Internet. Chapter 2 provides a review of the literature related to environmental scanning, the proposed methodology for solving the problem of developing an automated scanning process and the environment used to empirically test the methodology developed.

Chapter 3 describes the details of the methodology developed in this dissertation and the application environment used to empirically test the methodology. The methodology is tested by collecting news documents available via the Internet about publicly traded companies. Chapter 4 has additional details on the scanning process as well as a description of the experimental design used to empirically test the scanning process. The experimental design involves testing both a training set and a holdout sample for correct classification results. Chapter 5 presents the results. Finally, the Chapter 6 provides a summary, a conclusion and directions for future research.

CHAPTER 1
INTRODUCTION

Information is vital to an organization's survival. The amount of money spent on gathering intelligence by organizations is in the range of $50,000 to $1.5 million (Badeian 1986). Because of the Internet, the availability of useful information for a corporation is growing at a fast pace. The amount of information stored on the Internet has been estimated to double or more than double every 18 months (Yang et al. 2000). With all of this information available, companies must be able to obtain relevant documents and mine them for useful strategic decision-making information to gain competitive advantage.

Drucker (1998) claims that information technology (IT) has had little impact on making strategic decisions such as whether or not to enter a new market or build a new office building. According to a recent article by Denton (2001), organizations and individuals within the organization are drowning in too much information. However, Denton (2001) also claims that the Internet has the capability of improving knowledge management and should be used to simplify decision making by concentrating on critical information.

Scanning an organization's internal and external environment first became popular in the 1960s when scanning became a component of strategic planning (Russell and Prince 1992). Environmental scanning is still popular today. Ghoshal and Westney (1991) reported, based on a 1985 survey, that over one-third of a sample of *Fortune 500*

companies were spending over $1 million a year on competitive analysis. Additionally, Subramanian et al. (1993) reported that firms with environmental scanning systems to monitor the external environment had higher growth and profitability than firms that did not. General Electric, IBM, J. P. Morgan, Merrill Lynch, Motorola, Schlumberger and Xerox already use the Internet for information gathering, according to Pawar and Sharda (1997). Their article claimed that environmental scanning activities are in high demand. Additionally, they said that the availability of environmental scanning or business intelligence systems that meet company needs is inadequate.

Presently, one type of system used for scanning the environment, as well as other functions, is an executive information system (EIS). According to Moad (1988), many small firms and 70% of large firms currently use or are considering EIS. According to Bajwa et al. (1998), EIS is growing rapidly. Several EIS software packages are available including ACE Reports, COR Technology, e.Reporting Suite and eMis Executive Information Systems. However, past literature has reported failure rates between 40 and 70% for EIS (Raths 1989, Watson et al. 1991).

As the previous discussion suggests, there are several key reasons why an environmental scanning process for collecting, analyzing and interpreting information available on the Internet is desirable, including the importance of information to a corporation, the vast amount of information available via the Internet, the inability of many systems to process information for decision-making purposes, the link between profitability and external scanning systems and the lack of availability of scanning systems that meet companies' needs. Developing and testing a web-based scanning process is the essence of the research in this dissertation.

The process developed involves four main steps: (1) collecting web documents, (2) representing the documents collected via the popular vector space model (VSM) representation (Salton 1968), (3) separating a training set of documents using linear discriminant analysis (LDA) and (4) using the linear discriminant function determined by LDA to classify new documents according to their signal. Our process combines two well-established areas of research: the vector space model (VSM) for information retrieval developed by Salton (1968) and linear discriminant analysis first introduced by Fisher (1936) to scan web documents for signals that can aid in an organization's ability to make decisions.

1.1 Research Problem

As linear discriminant analysis is the technique used to classify documents according to their signals, one important problem is to determine how well LDA classifies the documents. Another problem is the degree to which new documents can be used for future classification. Both of these problems are investigated in this dissertation.

1.2 Purpose

The need for better tools for environmental scanning coupled with the availability of easily accessible information on the Internet presents a real need for an automated scanning tool that detects signals based on information gathered from the web. Additionally, the ability to use this tool to quantify and classify information for use by a business is important. Therefore, the purpose of this dissertation is to provide a framework for how a company can use the Internet to obtain external information about signals from the environment to help make better decisions. A process for automated web-based environmental scanning was developed and empirically tested in this study.

1.3 Organization

Chapter 2 provides a literature review of several areas related to the process developed in this dissertation. These areas are environmental scanning, the VSM for IR, linear discriminant analysis, equity investments in finance and text classification problems in computer science. Chapter 3 provides a detailed discussion of the problem setting, the application environment and research questions for this dissertation. Additionally, Chapter 3 provides the details of the VSM representation and the linear discriminant analysis problem. Chapter 4 has additional details on the environmental scanning process and discusses the experimental design. Chapter 5 provides the results of our empirical analysis of the process. Chapter 6 provides a summary of our findings, conclusions and directions for future research.

CHAPTER 2
LITERATURE REVIEW

In this chapter we review the literature related to this study of the automation of environmental scanning of the World Wide Web (WWW). Environmental scanning is the process of obtaining and using information from an organization's external environment to assist in decision-making (Aguilar 1967, Choo and Auster 1993). Lester and Waters (1989) said environmental scanning is a process that uses information from the environment to aid management in decision-making. There are three key components to this process. The first is obtaining the information, the second analyzing the information and the third using the information for making decisions in an organization (Lester and Waters 1989). According to Daft and Weick (1984), the way an organization deciphers its environment in order to learn from it may be divided into three phases: scanning, interpretation, and learning. Scanning involves information gathering. Interpretation involves giving meaning to the data. Learning involves taking action based on the data. All of these definitions have the common components of information gathering, analysis and use.

Scanning is a form of "organizational browsing." The benefits of scanning include finding the information desired by the organization, better decision making due to the use of information found during scanning, updating initial information requirements due to information discoveries, and the discovery of useful information found quite by accident. Scanning that is managed poorly can result in too much information that is

confusing. This can lead to wasting employees' time and energy, leading to high costs and wasted money with little or no action taken as a result of scanning (Choo 1995).

This study proposes methodology to automate the environmental scanning process through the use of web-based methods. The automated process combines two techniques: the vector space model for information retrieval introduced by Salton (1968) and discriminant analysis techniques originally introduced by Fisher (1936). The two techniques combined are used to solve a text classification problem. Additionally, the information being analyzed is news articles about stocks in the stock market. The chapter is organized as follows. Section 1 provides a literature review of environmental scanning and executive information systems (EIS). Section 2 is a review of the vector space model literature. Discriminant analysis is discussed in Section 3. Related work in finance is presented in Section 4. Finally, a brief overview of literature on the text classification problem is provided in Section 5.

2.1 Environmental Scanning

Environmental forces that affect an organization can be classified into six categories: demographics, economics, social and cultural, political and legal, technological, and competitive. Demographic forces involve the characteristics of the population being studied such as birthrate, family structure, divorce rate and education. Economic forces include inflation rate, spending patterns, income levels and purchasing power. Social and cultural forces that affect an organization involve the set of values, ideas and attitudes held by various cultures or social groups within the targeted consumer population. These attitudes can affect buying patterns of the population of consumers. Political and legal forces define the limits within which an organization can operate. The

federal government imposes regulations and policies in areas such as airline overbooking, food labels, prescription drugs, nursing homes and import tariffs (Michman 1983). Technological forces can affect an organization. Changing technology can dramatically change consumer markets (Michman 1983). Finally, competitive forces continually challenge organizations (Michman 1983). These forces are interrelated, yet beyond an organization's control.

Aguilar (1967) identifies four methods of environmental scanning: undirected viewing, conditioned viewing, informal search and formal search. In undirected viewing, scanning is done incidentally and without purpose. The manager is not prepared to analyze the information. Conditioned viewing involves exposure to selected information without searching for it. The manager is prepared to analyze or interpret the information. In informal search, information is actively sought without a structured search strategy. Formal search involves a manager actively seeking information using a formal search method.

Choo (1995) offered a different perspective on the types of scanning. Choo, like Aguilar, recognized four scanning methods. The first two, undirected and conditioned viewing, were first proposed by Aguilar (1967). The other two methods are enacting and discovery. Daft and Weick (1984) proposed that the type of scanning depends on how management perceives the analyzability of the external environment and depends on the intrusiveness of the organization. If management perceives the environment as highly analyzable, they will participate in more structured environmental scanning. Intrusiveness into the environment is determined by how willing an organization is to intrude or interfere with the external environment in order to understand it. An organization is

considered active or passive in terms of their organizational intrusiveness. An actively intrusive organization would seek out information while a passively intrusive organization would attempt to understand the external environment based on any information that happened to be provided without actively seeking that information out.

Based on the two dimensions described, organizational intrusiveness and analyzability of the environment, Choo (1995) proposed the scanning-interpretation model. According to the model, a passively intrusive organization that perceives the environment as unanalyzable will participate in undirected viewing. A passively intrusive organization that perceives the environment as analyzable will participate in conditioned viewing. An actively intrusive organization that perceives the environment as unanalyzable will participate in enacting. Organizations participating in enacting have a need to learn by doing, to do experimentation, often on the environment, and to seek information in the form of feedback from sources about the organization's activities. An actively intrusive organization that perceives the environment to be analyzable will participate in discovery. Discovery involves formal information needs that are satisfied through many sources through surveys and market research. "In summary, the scanning-interpretation model appears to be a viable framework for analyzing the primary environmental and organizational contingencies that influence environmental scanning as cycles of information seeking and information use activities." (Choo 1995, p.85)

Michman (1983) focuses on monitoring the changes that take place in the environment surrounding an organization and planning to make adjustments in the marketing strategy of an organization based on information obtained from monitoring.

Seven techniques are mentioned that were adapted to scanning the environment: trend extrapolation, Delphi technique, cross-impact analysis, simulation models, barometric forecasts, trend impact analysis, and multiple scenarios. Michman (1983) identified conditions, current in 1983, of demographic trends and their implications, the economic dimension of the environment, the social dimension, the political/legal dimension, the technological dimension, and the competitive dimension. A discussion of monitoring each of these dimensions and monitoring ecological changes is provided.

Jonsen (1986) considered environmental scanning by Universities. According to Jonsen (1986), environmental scanning is extremely important to the success of institutions of higher learning, and that importance is not yet realized. Although many university systems have environmental scanning structures already in place, these structures are not using all information available to the university. University systems typically scan for information about demographic and economic environments, ignoring the political, organizational (or competitive), technological and social aspects of the environment. Additionally, Jonsen (1986) proposes that even if universities have all six facets of information available, they do not have a system in place to understand the environment. They still need a way to understand and integrate the information. Universities need to be able to make decisions and plan based on the understanding and integration of the information. Also, universities need to make scanning of all facets of the environment, and using that information, a high priority.

Pawar and Sharda (1997) suggested using the Internet and its various utilities to scan the environment for external information. Their article classified the ability of the Internet to provide certain classes of information content. Additionally, the article

provided insight into the various types of search done over the Internet along with the Internet utilities best suited for those types of search. The types of search, originally discussed in Aguilar (1967), are undirected viewing, conditioned viewing, informal search and formal search. In the paper by Pawar and Sharda (1997) the various Internet utilities are classified as low, medium or high in terms of their ability to do the four different modes of search.

Camillus and Datta (1991) suggested that there are different patterns of acquiring information depending on the system of strategic planning used: strategic planning systems (SPS) and strategic issues management systems (SIMS). SPS focuses on information that is directly related to the organization and uses directed environmental scanning. SIMS monitors the environment continuously, looking for all signals, even signals that might be weak. Therefore, SIMS is more likely to use continuous undirected or semi-directed environmental scanning processes. Additionally, their article (Camillus and Datta 1991) suggested combining the systems for a semi-directed, continuous approach to environmental scanning.

Subramanian et al. (1993) conducted a survey to assess environmental scanning activities in Fortune 500 corporations. They used a method of classification from Jain (1984). Organizations can be classified over time as primitive, ad hoc, reactive or proactive in their scanning activities. The results based on the 101 firms that responded to the survey were 10% of firms operated in the primitive scanning mode, 30% in the ad hoc mode; 35% had reactive systems and 25% had proactive systems. Reactive and proactive scanning systems are considered advanced scanning activities.

Elofson et al. (1997) proposed a system architecture for sharing knowledge. The architecture consists of several nodes connected by a network. At each node there is a manager, a scheduler, a server, a planner, knowledge sources, one or more intelligent agents, and relevant data structures. The article discusses the use of multiple intelligent agents used to make decisions. The authors' framework provides for the distribution of knowledge across the organization, allowing for any division of the organization to gain access to the knowledge as needed. The decision support system architecture described in the article increases the ability of the organization to maintain a memory of, and to learn from, its decisions.

Walstrom and Wilson (1997) conducted a study based on information collected from 98 of the Corporate 1000 CEOs. The CEOs were asked how they used their EIS. The study was conducted to determine the types of EIS users and to classify how the different types used their EIS. Based on their study, they found three types of users and three underlying dimensions of usage. The three types of users were 'converts,' 'pacesetters' and 'analyzers.' 'Converts' use the EIS to increase their ability to access information. 'Pacesetters' use EIS to increase their ability to communicate and monitor performance. 'Analyzers' use EIS to solve problems. Walstrom and Wilson (1997) performed factor analysis to determine different types of uses of the EIS. The first, called 'organizational monitoring,' deals with using the EIS to monitor email, company news and organizational data. 'Information access' is the second dimension of use discovered by the authors. This second type is the most relevant use of the EIS to this paper. 'Information access' involves the use of the system to gain access to information both internal and external to the organization. The third dimension is called 'organizational

understanding' and involves using the system to help users understand the organization through ". . . unstructured access and analysis of organizational data" (Walstrom and Wilson 1997, p. 81). The authors then classify the value of each type of dimension to each identified type of user.

Koh and Watson (1998) conducted a study to examine data management issues in EIS. Based on their research a set of seven data management issues were identified. The authors then chose to study three in greater detail – data security, data ownership and data standards. They found that EIS users considered data standards to be the most important issue and also the most challenging. This is attributed to the fact that in EIS data come from different departments and levels of management. They tested several hypotheses, of which two were significant. There was correlation between the difficulty of data management and the breadth and depth of the data. Correlation also existed between the level of support from individuals and data management difficulty.

Bajwa et al. (1998) conducted a study to identify factors in the successful implementation of executive information systems. The study examined information system support, vendor/consultant support, and management support to determine their influence on the success of EIS. Their findings were based on data collected from sixty-nine firms. The authors found that IS support influences EIS success. No relation was found between management support or vendor/consultant support and EIS success. This finding is contrary to the literature. Additionally the authors found that management support influences IS support and vendor/consultant support.

In "Information Management for the Intelligent Organization," Choo (1995) conducted a review of pre-1995 literature on environmental scanning. The literature is

reviewed to study ". . . the effect of situational dimensions, organizational strategies, information needs, and personal traits on scanning behavior" (Choo 1995, p. 86). Choo (1995) grouped information needs into the following research categories: information needs as the focus of environmental scanning, information-seeking use and preferences, information seeking through scanning methods, and information use. The literature in each category was reviewed.

The first category of research, information needs as the focus of environmental scanning, focused on identifying the environmental sectors that are the primary focus of environmental scanning activity. Choo (1995) found that the literature (Aguilar 1967, Nishi et al. 1982, Ghoshal 1988, Johnson and Kuehn 1987, Lester and Waters 1989, Auster and Choo 1993a, Auster and Choo 1993b, Choo 1993, Olsen et al. 1994, Jain 1984) suggests that organizations consider information about customers, competitors and suppliers as most important.

The next category, information seeking use and preferences, relates to research that identifies and classifies sources of information. Information sources are classified as either internal or external to the organization and as personal or impersonal (Choo 1995). Personal sources convey information directly to the manager. Examples of impersonal sources are online databases, company libraries and publications (Choo 1995). Choo (1995) found that the literature (Aguilar 1967, Keegan 1967, Keegan 1974, O'Connell and Zimmerman 1979, Kobrin et al. 1980, Smeltzer et al. 1988, Lester and Waters 1989, Gates 1990, Mayberry 1991, Culnan 1983, Ghoshal and Kim 1986, Auster and Choo 1992, Auster and Choo 1993a, Auster and Choo 1993b) indicates that managers use both personal and impersonal sources as well as internal and external sources. Managers

prefer personal sources, especially when gathering information about customers, suppliers or competitors.

The third category of research, information-seeking scanning methods, was directed at classifying methods of environmental scanning and identifying the factors that affect the choice of scanning method (Choo 1995). Choo (1995) concluded that, based on the literature (Aguilar 1967, Keegan 1974, Thomas 1980, Klein and Linneman 1984, Preble et al. 1988, Subramanian et al. 1993, Wilson and Masser 1983, Al-Hamad 1988, McIntyre 1992, Fahey and King 1977), many factors affect the method of scanning an organization chooses to use including its size, industry category, environmental perception, dependence on the environment, experience in scanning, and experience with strategic planning.

The final category of research, information use, focuses on how the information obtained from scanning is used (Choo 1995). Many studies support that environmental scanning improves an organization's performance (Miller and Friesen 1977, Newgren et al. 1984, Dollinger 1984, West 1988, Daft et al. 1988, Subramanian et al. 1993, Subramanian et al. 1994, Murphy 1987, Ptaszynski 1989). Choo (1995) summarized that ". . . environmental scanning is increasingly being used to drive the strategic planning processes by business and public sector organizations in most developed countries." (Choo 1995, p.101) Additionally, there is a link between environmental scanning and how an organization performs.

2.2 Vector Space Model

Retrieving relevant information from a set of documents is not a new problem. As long as researchers have done research, techniques for finding information relevant to

that research have been developed and fine-tuned. Information retrieval is one of the critical areas in library science. Matching documents to users' queries has been the focus of much research. The problem of information retrieval still exists today, with a new twist. Modern information retrieval techniques involve searching the World Wide Web (WWW) for documents relevant to a user's query. The vector space model, one solution to retrieving relevant documents based on a user's query developed many years ago, is still applicable to today's information retrieval problem on the WWW.

In information retrieval, there is a set of documents and a user query. Based on the query issued, information retrieval schemes are used to return the most relevant documents from the collection to the user. Since relevancy is a judgment made by the user, information retrieval methods are not capable of producing only and all relevant documents. Therefore it is necessary to have a method of ranking the documents in terms of their similarity to the user's query. The vector space model (VSM) is a method of information retrieval that is used to retrieve and rank documents (Salton 1968).

The VSM, proposed by Salton (1968) involves representing a document by a vector of terms that are related to keywords or index terms in the document, as well as weights indicating the importance of these terms in indicating the content of the document. There are n index terms in a collection of k documents. The VSM notation is given in Table 1. Each index term corresponds to a vector of unit length. Let $t_1,\ldots,t_n$ be the vectors corresponding to the n index terms. In the basic model, the term vectors are assumed to be uncorrelated. This simplifying assumption results in pairwise orthogonality of the term vectors in the basic VSM. Document r is represented by vector,

d_r, such that each element of the document vector $d_{i,r}$ represents the weight of term i in document r.

$$d_r = \sum_{i=1}^{n} a_{i,r} t_i$$

When the terms are orthogonal, the elements of d_r are the term weights, which can be confirmed by $t_i' d_r = a_{i,r}$. For a collection of k documents, an $n \times k$ term by document matrix D can be constructed with the document vectors where each column of the document matrix corresponds to a document vector d_r.

Table 1. Vector space model (VSM) notation

Variable	Meaning
k	Number of documents
n	Number of index terms
t_i	$n \times 1$ term vector representing term i
T	$n \times n$ term matrix where t_i's are the columns
d_r	$n \times 1$ document vector of n index terms
D	$n \times k$ matrix where d_r's are the columns
$a_{i,r}$	represents the weight of term i in document r
q	$n \times 1$ query vector where q_i represents the weight of term i in query q

In the VSM (Salton 1989), a vector $q' = (q_1, \ldots, q_n)$ is used to represent a user query. The vector q is then matched against the document vectors to determine which documents are most similar to the query. To rank the documents according to similarity, a similarity measure is used. Several similarity measures have been suggested, including the commonly used scalar product between the query vector q and document vectors. The scalar product between two vectors x and y is defined as

$$x'y = |x||y|\cos\theta$$

where $|x|$ is the length of the vector x and θ is the angle between the vectors x and y.

The document-query similarity can be computed by

$$d_r' q = \sum_{i,j=1}^{n} a_{i,r} q_j t_i' t_j .$$

The term-term correlation, $t_i \cdot t_j$, is not usually known a priori and in the basic model, the terms are assumed to be uncorrelated and hence, orthogonal. Therefore, in the basic model the term-term correlation is given by

$$t_i \cdot t_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}.$$

This results in the orthogonality of the term vectors, creating a linear independent set of n vectors. The term vectors $t_1, \ldots, t_n$ are therefore a basis for the document space. When the term vectors are assumed to be uncorrelated, the document-term similarity is reduced to

$$d_r' q = \sum_{i=1}^{n} a_{i,r} q_i .$$

An example of the basic model is provided in Appendix A.

According to Salton (1989), document-document similarity can be measured using the same concept as document-query similarity. The similarity between two documents is given by

$$d_r' d_s = \sum_{i,j=1}^{n} a_{i,r} a_{j,s} t_i' t_j$$

which reduces to

$$d_r' d_s = \sum_{i,j=1}^{n} a_{i,r} a_{j,s}$$

when the term vectors are uncorrelated, hence orthogonal. Computing document-document similarity is useful for determining how to organize the documents in the document space.

Table 2. Measures of query-document similarity

Inner product	$d_r q = \sum_{i=1}^{n} a_{i,r} q_i$
Cosine	$d_r q = \frac{\sum_{i=1}^{n} a_{i,r} q_i}{\sqrt{\sum_{i=1}^{n} q_i^2 \sum_{i=1}^{n} a_{i,r}^2}}$
Pseudo-cosine	$d_r q = \frac{\sum_{i=1}^{n} a_{i,r} q_i}{\sum_{i=1}^{n} q_i \sum_{i=1}^{n} a_{i,r}}$
Dice	$d_r q = \frac{2\sum_{i=1}^{n} a_{i,r} q_i}{\sum_{i=1}^{n} q_i + \sum_{i=1}^{n} a_{i,r}}$
Covariance	$d_r q = \sum_{i=1}^{n} (q_i - \bar{q})(a_{i,r} - \bar{d}_r)$
Product-moment correlation	$d_r q = \frac{\sum_{i=1}^{n} (q_i - \bar{q})(a_{i,r} - \bar{d}_r)}{\sqrt{\sum_{i=1}^{n} (q_i - \bar{q})^2 \sum_{i=1}^{n} (a_{i,r} - \bar{d}_r)^2}}$
Jaccard Coefficient	$d_r q = \frac{\sum_{i=1}^{n} a_{i,r} q_i}{\sum_{i=1}^{n} q_i^2 + \sum_{i=1}^{n} a_{i,r}^2 - \sum_{i=1}^{n} a_{i,r} q_i}$
Overlap	$d_r q = \frac{\sum_{i=1}^{n} \min(q_i a_{i,r})}{\min\left(\sum_{i=1}^{n} q_i, \sum_{i=1}^{n} a_{i,r}\right)}$

As previously mentioned, several methods of measuring the similarity between a document d_r and a query have been suggested. A list of these measures is provided in

Table 2 under the assumption that the term vectors are orthogonal where $\bar{q} = \frac{1}{n}\sum_{i=1}^{n} q_i$ and $\bar{d}_r = \frac{1}{n}\sum_{i=1}^{n} a_{i,r}$. The cosine, Dice, Jaccard, pseudo-cosine, and product-moment correlation measures are all normalized to fall within the range [0, 1] for nonnegative vector elements.

An interpretation of several of the measures is given in Jones and Furnas (1987). The inner product measures similarity as a simple weighted sum. When the term weights in both the document and query vector are binary, the inner product measure is the cardinality of the intersection of the two vectors q and d_r. The cosine measure is the inner product measure with each of the document and query vectors normalized by their Euclidean or ℓ_2 lengths. The normalization provides that the cosine measure ranges from 0 (no matching terms) to 1 (perfect match between the vectors). The pseudo-cosine measure is the inner product measure with the document and query vectors normalized by the ℓ_1 or city-block lengths. The Dice measure is twice the inner product measure divided by the sum of the ℓ_1 lengths of the vectors. The covariance measure is the inner product measure of two new vectors $q' = (q - \bar{q})$ and $d_r' = (d_r - \bar{d}_r)$ where q' is the average term weight in the query vector, q, subtracted from q and d_r' is the average term weight in the document vector, d_r, subtracted from d_r. The product-moment correlation measure is the cosine measure of the two new vectors q' and d_r'. The overlap measure, like the cosine measure, ranges from 0 to 1. The numerator of the overlap measure sums, for each component of the query and document vectors, the minimum term weight values. When the vectors are binary, the numerator of the overlap

measure and the inner product measure are the same. The denominator is the minimum of the ℓ_1 lengths of the two vectors. When the term weights are binary, the Jaccard measure can simply be thought of as the size of the intersection of the query and document vectors divided by the size of the union of the two vectors.

Wang et al. (1992) show that when several of these measures are normalized they are linear by the general definition provided in their paper. The definition is given below.

> Let $m(q,d)$ be a similarity measure on $\Re^n \times \Re^n$. If there exists two functions:
>
> $$N_q : \Re^n \to \Re^n,\ q \to \hat{q} = N_q(q)$$
>
> and
>
> $$N_d : \Re^n \to \Re^n,\ d \to \hat{d} = N_d(d)$$
>
> such that
>
> $$m(q,d) = N_q(q) \cdot N_d(d) = \hat{q} \cdot \hat{d},$$
>
> then we say $m(q,d)$ is a linear measure. The function N_q can be regarded as a normalization of the query vector and the function N_d as a normalization of the document vectors. (Wang et al 1992, pp. 154)

Using the above definition, Wang et al. (1992) demonstrate the linearity of the inner-product, cosine, pseudo-cosine, covariance and product-moment correlation measures are linear. The Dice measure is shown to be linear under the special case that the query vector $q = (q_1, \ldots, q_n)$ is normalized with the ℓ_1 norm, i.e., $|q| = \sum_{i=1}^{n} q_i = 1$. The authors found the necessary and sufficient conditions under which the linear measures of similarity between documents and queries would produce an acceptable ranking of the documents. "The important point is that this result establishes the basis for adopting these similarity measures in information retrieval." (Wang et al 1992, pp. 159)

Indexing algorithms are used to determine the set of index terms in the vector space model. Indexing can be very simple or quite complex. In Salton (1989) several methodologies for indexing are provided. The indexing methodologies involve several steps, including a step to compute the weight of term, t_j, for document d_i. The term weighting schemes involve computing term frequencies, inverse document frequencies, term discrimination values and term-relevance weights.

The term frequency, $tf_{i,j}$, is the frequency of word stem t_j in document d_i. The document frequency, df_j, is the number of documents in the collection that contain term t_j. The inverse document frequency (idf) is given by $\log\left(k/df_i\right)$, where k is the number of documents in the collection.

In Salton (1989), the term discrimination value, dv_j, for term t_j is given by the difference between the space density of the document collection without the term, Q, and the space density with the term, Q_j.

$$dv_j = Q - Q_j$$

The space density of a document collection is ". . . the average pairwise similarity between all pairs of distinct items." (Salton 1989, p. 282)

$$Q = \frac{1}{k(k-1)} \sum_{m=1}^{k} \sum_{\substack{n=1 \\ m \neq n}}^{k} \mathrm{sim}(d_m, d_n) \qquad (1)$$

The higher the average pairwise similarity between the distinct documents the denser the document space. The addition of a term t_j with good discriminatory power will result in a document space that is less dense. Hence, terms that are good discriminators have

positive term discrimination values. Computing the space density of the document collection with (1) is quite expensive computationally. A more efficient computation involves defining a document centroid, $C = (c_1, c_2, \ldots, c_n)$, at the center of the document space. Element j of the centroid is defined as the average value of the j^{th} terms in the document collection, d_{ij}, and is given by

$$c_j = \frac{1}{k}\sum_{i=1}^{k} d_{ij}.$$

The space density is the average similarity between each document in the collection and the centroid.

$$Q = \frac{1}{k}\sum_{i=1}^{k} \mathrm{sim}(c, d_i)$$

This measurement involves calculating k similarities as opposed to $k(k-1)$ similarities in (1).

According to Salton (1989), the retrieval value of each term t_j in a document is given by

$$tr_j = \log \frac{p_j(1-q_j)}{q_j(1-p_j)}$$

and is called the term-relevance weight. The term-relevance weight is determined by the probability of occurrence of term t_j in π_1 or π_2. Let r_j be the number of documents in π_1 that contain the term t_j. Assuming that there are k_1 (k_2) documents in π_1 (π_2), then

$$p_j = \frac{r_j}{k_1}$$

and

$$q_j = \frac{df_j - r_j}{k_2}$$

where p_j is the probability of term t_j occurring in π_1 and q_j is the probability of term t_j occurring in π_2.

A single term indexing algorithm using the term frequency multiplied by the inverse document frequency, the term discrimination value or the term-relevance weight follows as elements of the document vector.

1. After cleaning the documents, list every word in the document collection.
2. Run a *suffix-stripping* routine to convert all words to their word stems.
3. Sort the list of word stems and write as a vector v.
4. For every word stem, t_j, remaining in the document d_i, compute a weighting factor, w_{ij}, the weight of term t_j in document d_i.
   a) The weighting factor can be computed by $w_{ij} = tf_{i,j} \cdot \log\left(k/df_i\right)$ which is composed of the term frequency and inverse document frequency factor.
   b) The weighting factor can be computed by $w_{ij} = tf_{i,j} \cdot dv_j$ which is composed of the term frequency and the discrimination value of term, t_j, for document d_i.
   c) The weighting factor can be computed by $w_{ij} = tf_{i,j} \cdot tr_j$ which is composed of the term frequency and the term-relevance weight of term, t_j, for document d_i.
5. Eliminate all word stems in the vector v with $w_{ij} \leq q$, where q is a chosen threshold value.
6. For each document d_i in the collection, construct the term weight vector a_i as the weights of each word stem in the place where the stem occurs in v and a zero in all other places.

One of the problems inherent in the basic VSM is that there are no guidelines in choosing the similarity measure and that choice is left to the user (Salton 1989). Wong et

al. (1988, Bollmann and Wong 1987, Wong and Yao 1990) address this problem by developing ranking strategies based on user preference instead of the idea of relevance of documents. Two ranking strategies are developed called the perfect and acceptable rankings. These ranking strategies are found to have a great impact on information retrieval strategies. Wang et al. (1992) develop the necessary and sufficient conditions for linear similarity measures used for ranking documents to produce an acceptable ranking strategy. The authors also examine the geometric properties of the various linear similarity measures and analyze the structure of the solution vectors. The work in Wang et al. (1992) extends the work in Wong et al. (1988, Bollmann and Wong 1987, Wong and Yao 1990).

As with any model, measures are needed to determine the performance of the model. Recall and precision are measures that are often used to determine the performance of information retrieval methods. Documents can be partitioned in a document collection into the set of documents relevant to the query, A, and the set not relevant to the query, $\overline{A}$ (Salton 1968). Additionally, documents can be partitioned into the set retrieved by the system according to the query, B, and the set not retrieved, $\overline{B}$. Recall is defined as the ratio of the number of documents retrieved that are relevant to the user's query to the total number of documents that are relevant in the document collection. Assume $|A|$ is a count of the number of members in set A. Recall is given by

$$\text{Recall} = \frac{|A \cap B|}{|A|}.$$

Precision is the ratio of the number of documents retrieved that are relevant to the user's query to the total number of documents retrieved. Precision is given by

$$\text{Precision} = \frac{|A \cap B|}{|B|}.$$

There exists a tradeoff between the level of recall achieved and the level of precision.

Salton et al. (1975) describe using the vector space model to classify documents based on their similarity of terms. They determine how to best configure the document space through automatic indexing. Ideally, documents that are similar in index terms according to some similarity measure will be close spatially and documents deemed dissimilar are not spatially close to one another. Their paper examines the relationship between the density of the document space and the performance of indexing techniques.

The paper by Salton et al. (1975) describes two methods of configuring the document space. The first involves clustering documents based on whether a given set of documents is often used simultaneously in response to a user's query. If one document in a cluster is deemed similar to a particular query, the "neighbors" in the cluster should also be returned in response to the query. The problem with the clustering method of document space configuration is that the retrieval history for the document collection must be known in order to ascertain the user's input about the relevance of the documents in relation to the query. The next best document indexing method suggested by Salton et al. (1975) is to maximize the space between the set of documents. This is achieved by minimizing the sum of all similarity measures between distinct documents over the entire set of documents. The problem with this approach is that the order of complexity of the solution is n^2. Although this is polynomial, n is usually very large and, hence, a

clustered document space is used in all further studies on the document space and indexing.

Salton et al. (1975) examine the correlation between performance of the indexing technique and the density of the document space. Based on their study, lower document space density seems to result in better recall-precision performance. Continuing the exploration of the relationship between density and performance, Salton et al. (1975) examine the effects of changing the document space to determine whether the changes cause differences in the recall-precision performance. This is done by increasing the similarity within clusters and decreasing the between cluster similarity. The results again indicate that dense document spaces correspond to lower recall and precision levels.

Salton et al. (1975) also discuss a discrimination value model (DVM) first introduced by Salton and Yang (1973) and Salton (1975). The value of an index term is based on its ability to discriminate among documents by increasing the difference among document vectors when the term is assigned as an index term in the document collection (Salton and Yang 1973, Salton 1975).

The SMART retrieval system is a document retrieval system used as an experimental tool for evaluating various search procedures (Salton 1971). The system was designed at Harvard between 1961 and 1964 and was fully implemented at both Harvard and Cornell. The SMART system at Cornell consists of five sections: (1) text input, (2) clustering or grouping documents, (3) selection of document groups for search, (4) searching and (5) evaluation (Williamson et al. 1971). A simplified SMART system flowchart is provided in Salton and McGill (1983, p. 129). Experiments with the SMART system have been done on several document collections: IRE-3 abstracts in

computer literature, CRAN-1 a collection of documents in aerodynamics, Ispra documents in documentation, ADI short papers in documentation and Medlars a collection of documents in medicine (Salton 1971).

The results of the experiments are summarized by Salton (1971) according to document length, term weights and matching functions, word normalization, dictionaries with synonym recognition, phrase generation methods, hierarchical procedures, fully automatic versus manual text processing, feedback procedures and partial cluster searches. The longer the length of the text examined to match queries to documents, the better the retrieval performance. However, the improvement in performance is not enough to conclude that a full text search is always warranted over searching abstracts only. Weighted terms are more effective for describing the content of a document than terms without weights. The cosine measure of similarity between a document, d_i, and a query, q, is more useful than an overlap function. Further improvements can be made by using more complex measures of similarity and synonym recognition. Word normalization is most useful when documents contain non-technical, redundant vocabulary. Phrase generation methods and hierarchical procedures are not of sufficient use, in general, to merit implementing them in automatic systems. User feedback procedures substantially improve performance of subsequent iterations of search. Search time can be reduced greatly by using partial cluster searches of selected groupings of documents as opposed to a complete match of the query with all documents stored. Manual-indexing methods were not found to be substantially superior to fully automatic text processing. Finally, Salton (1971) ranks the tested procedures on their order of merit as follows, with the most effective listed first:

> (1) abstract processing with phrase and synonym recognition, (2) weighted word stem matching and statistical word associations using abstracts for analysis purposes, (3) logical word stem matching disregarding term weights and (4) title processing using only document titles for analysis purposes, and document-request matching based on overlap function. (Salton 1967, p. 6)

According to Raghavan and Wong (1986), in order to use the VSM, the various components of the model must be specified. The dimension of the space is assumed to be n , the number of distinct terms. This need not be the case. The term-term correlations need to be provided. The simplifying assumption in the literature is that the term vectors are pairwise orthogonal. Additionally, an interpretation of the elements of D needs to be given. Typically, the elements of D are interpreted as the components of document vectors along the direction of the term vector (Salton 1989).

Raghavan and Wong (1986) present the vector space model relaxing the assumption of term vectors being pairwise orthogonal. In this case, they illustrate the difference between components of documents along terms and projections of documents along terms. They also discuss the standard vector space model where term vectors are pairwise orthogonal. They illustrate that in this model the elements of D are both projections and components of documents along term vectors. The authors discuss the dual of the standard vector space model, where the elements of D are interpreted as components of terms along documents, as opposed to documents along terms. A case is made for the use of negative term correlations in the model, where a negative correlation between terms would indicate the degree to which the terms are "opposite".

Raghavan and Wong (1986) present several shortcomings and misconceptions involving the VSM to date along with clarifications in the interpretation of the components of the VSM. One of the points made is that the linear dependence of term

vectors, orthogonality and dimensionality of the vector space are concepts that are closely related. If all of the term vectors are uncorrelated, meaning the terms are pairwise orthogonal, then the set of term vectors is linearly independent. However, the converse is not necessarily true. Linear independence of the term vectors does not imply uncorrelated terms. "Rather, linear independence only implies that any redundancy in the usage of terms has been removed and the representation in terms of the resulting set of vectors is compact (and unique)." (Raghavan and Wong 1986, pp. 286) In much of the earlier literature the distinction between linear independence and non-orthogonality is not present.

Vector space representations that use linearly dependent term vectors may have terms that can be removed without losing information (Raghavan and Wong 1986). This idea stems from the linear algebra concept that if a vector space is spanned by a linearly dependent set of vectors, then the vector space can be spanned by a linearly independent subset of the vectors.

Raghavan and Wong (1986) demonstrate that choosing to interpret the elements of D as ". . . both the component of documents along term vectors as well as the components of terms along document vectors is inconsistent." (Raghavan and Wong 1986, p. 287) To illustrate why this interpretation is inconsistent, the authors discuss the standard VSM, where D is interpreted as the component of documents along term vectors, $D = A$, compared to the dual of the standard VSM, where D is interpreted as components of terms along document vectors, $D = B'$. The problem that arises when both interpretations are used, as is sometimes done in the literature according to the authors, is that $A = B'$. This is only true in the special case where both the term-term

correlation matrix and the document-document correlation matrix are identity matrices. As the authors point out, this case is uninteresting.

Several papers have been written to expand the vector space model in the face of several shortcomings presented by the basic model. In Wang et al. (1992), necessary and sufficient conditions are identified for ranking documents under the various similarity measures. In order to relax the assumption of pairwise orthogonality of the term vectors, correlation between term vectors must be known (Wong et al. 1987, Raghavan and Wong 1986). Crouch, Crouch and Nareddy (1990) provide a method of automatically constructing an extended query containing multiple concept classes based on a query issued with only a single term. Kleinberg and Tomkins (1999) discuss how to deal with the problem of the high dimensionality inherent in the VSM by a dimension reduction technique, Latent Semantic Indexing. The idea of Latent Semantic Indexing was originally introduced in Deerwester et al. (1990). Additionally, the authors contrast this technique with other approaches to the high dimensionality problem, including clustering and vector space dimension-reduction. The authors also provide a discussion on documents on the World Wide Web, where there is a link-structure present. Henrich (1996) discusses spatial access methods and their application to the document vectors in the VSM. Henrich's paper (1996) also has to deal with the problem of dimensionality. Spatial access methods, such as nearest neighbor and distance scan queries, assume a small dimension, while the VSM typically involves very high dimensionality.

Wong et al. (1987) developed a generalized vector space model (GVSM). The method of computing term-term similarity is based on term co-occurrence using three ideas. First, a term or concept is defined by the documents that contain that concept. A

term is unrelated to another term if the set of documents containing the term does not intersect with the set containing the other term. Finally, the more documents in the intersection between two terms, the greater the similarity measure between them. The GVSM is presented and theoretically justified based on the previously stated ideas behind measuring term-term similarity.

Table 3. Variables for two-group linear discriminant analysis

Variable	Description
p	Number of attributes
x	Column vector $p \times 1$ of attributes
N	Number of observations, $N = N_1 + N_2$
N_i	Nonzero number of observations from group i, $i = 1, 2$
X_i	$N_i \times p$ matrix whose jth row is the transpose of the jth observation vector from group i, $i = 1, 2$, $j = 1, \ldots, N_i$
π_i	Group/class i, $i = 1, 2$
z_i	Scalar cutting score for group i classification, $i = 1, 2$
w	Column vector $p \times 1$ of coefficients determined by discriminant analysis
1_i	Column vector $N_i \times 1$ of ones, $i = 1, 2$

2.3 Discriminant Analysis

Discriminant analysis is a technique used to separate observations into classes or to assign new observations to already existing classes based on a discriminant function or discriminant score (Johnson and Wichern 1982). In it's linear form, discriminant analysis separates observations via one or more hyper-planes. The general objective of discriminant analysis is to find a hyper-plane that best separates the observations. The method for finding the hyper-plane, the specific objective, the criteria used to find cutting score z and the underlying assumptions are different for each variation of the discriminant analysis methods. "All of the models determine a linear discriminant function (LDF) by optimizing some criterion that is invariably a surrogate for minimizing

the number of misclassifications." (Koehler and Erenguc 1990, p. 63) An observation x is misclassified if the discriminant function places x in the wrong group (Koehler and Erenguc 1990).

First, a discussion of the notation used in this section is provided in Section 2.3.1. Second, a description of Fisher's (1936) linear discriminant function is given in Section 2.3.2. Finally, several linear programming variations are presented in Section 2.3.3.

2.3.1 Notation

For all variations of two-group discriminant analysis, the discriminant function is found based on observations for which the group classification is known. This group of observations is called the training sample (Koehler and Erenguc 1990). Adopting the notation from Koehler and Erenguc (1990) and Johnson and Wichern (1982), Table 3 describes the variables from the observations in the training sample. The classification of a column vector of observations x is assigned to π_1 if $w'x \geq z_1$ and π_2 if $w'x \leq z_2$. In most variation z_1 and z_2 are equal. In the case where they are equal, observations that lie on the hyperplane are arbitrarily classified into π_1 or π_2.

2.3.2 Fisher's Linear Discriminant Function

Fisher originally proposed discriminant analysis in a paper in 1936 (Fisher 1936). The assumptions in Fisher's (1936) linear discriminant analysis method are (1) the data are multivariate, normally distributed and (2) the variance-covariance matrices of the two groups of data are known and equal. Fisher (1936) decided to use the linear combination of x that maximized the ratio of the distance between the means to the variance in the Y 's given by:

$$\frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2},$$

where μ_{1Y} is the mean of the Y 's obtained from X 's belonging to π_1, μ_{2Y} is the mean of the Y 's obtained from X 's belonging to π_2 and σ_Y^2 is the variance in the Y 's. Fisher's linear discriminant function is the maximum of the above ratio. The function is given by

$$(\mu_1 - \mu_2)' \Sigma^{-1} X .$$

The covariance matrices are assumed to be the equal for both classes of objects. The covariance matrix is given by

$$\Sigma = E(X - \mu_i)(X - \mu_i)', \qquad i = 1, 2 .$$

When μ_1, μ_2, and Σ are unknown Fisher's sample linear discriminant function is used. The function is given by

$$y = (\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x$$

$$S_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}.$$

The cutting score z depends on a point between $\bar{y}_1 = b'\bar{x}_1$ and $\bar{y}_2 = b'\bar{x}_2$. If the group sizes of the two classes are equal, use the midpoint to determine the cutting score z. If the group sizes are unequal, use a weighted average method.

2.3.3 Linear Programming Approach to Linear Discriminant Analysis

Mangasarian (1965) and Rosen (1965) proposed mathematical programming methods for discriminant analysis. Later, Hand (1981) and Freed and Glover (1981a, 1981b) independently examined the linear programming approach to the discriminant

problem in more depth. The advantage of using a linear programming approach is the lack of assumptions that are present in Fisher's LDF. Various linear programming approaches show promise, especially when the assumptions for Fisher's LDF are violated (Koehler 1989a, 1989b, Ragsdale and Stam 1991). Freed and Glover (1981a) proposed two LP formulations of the discriminant analysis problem, the minimize the maximum deviations (MMD) model and the minimize the sum of deviations (MSD) model. Hand (1981) proposed the perceptron model, a version of an MSD model that appeared independently and in the same year as the MSD model proposed by Freed and Glover (1981a). Freed and Glover (1981b) proposed another model called the minimize the sum of interior distances (MSID) model.

The MMD formulation is discussed in Section 2.3.3.1. The MSD formulation is discussed in Section 2.3.3.2. In Section 2.3.3.3 a discussion of problems that occur in the formulations is given. Finally, in Section 2.3.3.4 some alternative formulations are discussed to alleviate the problems discussed in Section 2.3.3.3.

2.3.3.1 The MMD formulation

The MMD formulation (Freed and Glover 1981a) the objective is to minimize the maximal violations of misclassified observations. The original MMD formulation (Freed and Glover 1981a) is

$$\text{Maximize } d$$

Subject to

$$X_1 w - dl_1 \geq cl_1$$

$$X_2 w + dl_2 \leq cl_2$$

$$w, d \qquad \text{unrestricted in sign}$$

c positive constant

If $d > 0$, then perfect group separation is achieved. If $d < 0$, then there is some overlap in the groups, but the overlap is minimized. If $d = 0$, then the two groups share the points on the hyperplane. Several variations of the MMD formulations are provided in Freed and Glover (1986a,b).

2.3.3.2 The MSD formulation

The MSD formulation (Freed and Glover 1981a) minimizes the sum of the (weighted) violations. One of the original MSD formulations given by Freed and Glover (1981a) is

$$\text{Maximize } l_1' d_1 + l_2' d_2$$

Subject to

$$X_1 w - d_1 \geq cl_1$$

$$X_2 w + d_2 \leq cl_2$$

w, d_1, d_2 unrestricted in sign

c positive constant

The original MSD formulation proposed independently by Hand (1981) is

$$\text{Minimize } l_1' d_1 + l_2' d_2$$

Subject to

$$X_1 w + cl_1 + d_1 \geq b_1$$

$$X_2 w + cl_2 - d_2 \leq -b_2$$

$$d_1, d_2 \geq 0$$

w, c unrestricted in sign

b_1, b_2 positive constants

Other MSD formulations are given in Freed and Glover (1986a) and Bajgier and Hill (1982).

2.3.3.3 Discussion of problems with formulations

There are problems with some of the LP approaches: unbounded solutions, improper solutions, unacceptable solutions and instability of solutions when performing a linear transformation of the data (Ragsdale and Stam 1991). A solution is said to be unbounded in LP if the objective function can be increased or decreased without limit. An improper solution occurs when all observations lie on the classification hyper-plane (Ragsdale and Stam 1991). If a solution ". . . generates a discriminant function of zeros, in which case all observations will be classified in the same group . . ." then it is an unacceptable solution. (Koehler 1989a, p. 241) Koehler (1989a) characterizes unacceptable solutions for the various versions of the MMD, MSD and MSID models. Some LP versions of discriminant analysis have provided different solutions to the LP problem under linear transformation of the data (Markowski and Markowski 1985). When this problem occurs, the solution is instable under linear transformation.

Glorfield and Gaither (1982) criticize the LP approach offered by Freed and Glover (1981a). The criticisms involve the simplicity of the LP model, the appropriate role of the LP model, the adaptability of the LP model to more than the two-group problem and ability of the LP model to compete with existing approaches. Freed and Glover answer the criticisms in (Freed and Glover 1982).

2.3.3.4 Alternative formulations

Several alternative LP formulations of the discriminant problem have been offered in response to the problems discussed in 2.3.3.3. In an attempt to avoid unacceptable solutions, Glover et al. (1988) suggest a hybrid formulation called the HDM. Upon the discovery by Koehler (1991) that the formulation did not completely solve the problem of unacceptable solutions, Glover (1990) refined the HDM. The refined version does solve the problems that plagued earlier versions.

However, due to the increased complexity of the refined version of the HDM, Ragsdale and Stam (1991) offer simplified formulations of the earlier MMD and MSD models based on using a classification gap. The formulation based on the MMD is called the epsilon MMD or the EMMD and is given by

$$\text{Minimize } d$$

subject to

$$w_0 l_1 + X_1 w - d l_1 \leq 0$$

$$w_0 l_2 + X_2 w + d l_2 \geq \varepsilon l_2$$

$$d \geq 0$$

$$w_0, w \qquad \text{unrestricted in sign}$$

where w_0 is the intercept and $\varepsilon = 1$ is the gap between two hyper-planes found as a solution. The epsilon MSD (EMSD) formulation developed in Ragsdale and Stam (1991) is given by

$$\text{Minimize } 1_1' d_1 + 1_2' d_2$$

subject to

$$w_0 l_1 + X_1 w - d_1 \leq 0$$

$$w_0 l_2 + X_2 w + d_2 \geq \varepsilon l_2$$

$$d_1, d_2 \geq 0$$

w_0, w unrestricted in sign.

Additionally, Koehler and Erenguc (1990) provide a mixed integer mathematical programming formulation to minimize the number of misclassifications.

2.4 Financial Literature

There are several recent papers investigating the relationship between news and stock returns. Chan (in press) studies the return patterns for a set of stocks with public news releases (news stocks) versus a set of stocks with similar monthly returns without news releases (no-news stocks). He finds that there is a major difference in the return patterns for news versus no-news stocks. Specifically, news stocks with negative returns underperformed their peers, positive news stocks experience less price drift and extreme return no-news stocks experience reversal in the subsequent month with little abnormality after that.

In a recent study by Daniel and Titman (Daniel, K., & Titman, S. 2001. Market reactions to tangible and intangible information. Unpublished manuscript.), reactions to tangible and intangible information in the market are examined.

> Tangible information consists of explicit performance measures, like sales, earnings and cash flows, which can be observed in the firms' accounting statements. Intangible information, in contrast, is that part of the stock's past return that cannot be linked directly to accounting numbers, but which presumably reflects expectations about future cash flows. (Daniel, K., & Titman, S. 2001. Market reactions to tangible and intangible information. Unpublished manuscript, p. 3)

They find evidence that investors overreact to intangible information, but not to tangible information. To explain these results, Daniel and Titman (Daniel, K., & Titman, S. 2001. Market reactions to tangible and intangible information. Unpublished manuscript.) highlight work in psychology that suggests individuals are overconfident when interpreting information in settings where more judgment is required and short-run feedback on the value of the judgment is unavailable (Einhorn 1980).

In a paper by Huberman and Regev (2001), the events surrounding an in-depth news story in *The New York Times* on Sunday, May 3 1998 concerning a breakthrough in cancer research by Entremed Inc. (ENMD) are examined. After the story, the stock price of ENMD increased by 330% from Friday-close to Monday-close and the price of stocks in the biotechnology sector also increased significantly. Interestingly, the scientific breakthrough discussed in the story had already been published in *Nature* and in other sources in the press five months prior to the article in the *Times*, with a much milder reaction with respect to stock price and no spillover to the rest of the biotechnology sector.

2.5 Text Classification

The problem of analyzing the content of text-based documents for classification purposes is not new. The purpose of text categorization is to classify each document in a document collection into zero to multiple categories based on a predefined set of categories. Many algorithms have been written for the text classification problem including CONSTRUE (Hayes & Weinstein 1990), DTree (Lewis & Ringuette 1994), NaiveBayes (Lewis & Ringuette 1994), SWAP-1 (Apte et al. 1994), Nnets (Wiener et al. 1995), Rocchio (Rocchio 1971), k-NN (Hayes & Weinstein 1990) and support vector

machines (SVM) introduced by Vapnik et al. (Vapnik 1995, Cortes & Vapnik 1995). Many of the algorithms are based on statistical learning methods and some are based on the VSM (Salton 1968). The process discussed in this dissertation is based on the VSM (Salton 1968) in accordance with the most popular text categorization algorithm, Rocchio (Rocchio 1971) also based on the VSM.

In Rocchio (1971), each document vector is categorized according to a set of prototype vectors. Each predefined category has a prototype vector constructed based on a training set of documents. Each document is ranked according to a similarity measure comparison between the document vector and each prototype category vector. CONSTUE (Hayes & Weinstein 1990) is a rule-based expert system used to categorize Reuters news stories. DTree (Lewis & Ringuette 1994) methods of classification are based on decision trees. NaiveBayes (Lewis & Ringuette 1994) is a text classification method based on a naïve Bayes model. SWAP-1 (Apte et al. 1994) uses rules with an inductive learning algorithm for classification. Wiener et al. (1995) use neural networks (NNets) to solve the text classification problem. Hayes and Weinstein (1990) developed a k-nearest neighbor (k-NN) approach to solving the classification problem. Yang (1999) evaluates the aforementioned methods of text categorization on two document collections, the Reuters corpus, newswire stories collected from 1987 to 1991 and the OHSMED corpus developed at the Oregon Health Sciences University by William Hersh and colleagues. Yang (1999) finds that of the aforementioned methods, k-NN and NNets were the top performers. Support vector machines were introduced by Vapnik et al. (Vapnik 1995, Cortes & Vapnik 1995) and are based on statistical learning theory. Joachims (1998) compares the performance of support vector machines to the Rocchio

algorithm, the naïve Bayes classifier, the k-NN classifier and the C4.5 decision-tree rule learner (Quinlan 1993). Joachims (1998) concludes that support vector machines outperform the methods to which they were compared significantly.

The first problem encountered in text categorization is how to represent the documents. Salton (1968) developed the VSM to represent documents and queries issued by users as vectors. Using vectors to represent documents provides a quantitative approach to the problems of information retrieval and text categorization. The basic idea in the VSM is to convert documents to vectors by first converting each word in the document to it's word stem and then constructing the document vector by counting the frequency of each word stem in that document.

After a suitable document representation has been decided, the next problem in text classification is the method of categorization. As the VSM provides a linear representation of documents, it is natural to categorize the documents via linear discriminant analysis (LDA) or natural generalizations such as support vector machines (Vapnik 1995, Cortes & Vapnik 1995). This is in contrast to the categorization method used in the Rocchio (Lewis et al. 1996) algorithm.

CHAPTER 3
RESEARCH PROBLEM

As discussed in Section 2.1, environmental scanning is very important to the success of a corporation. Many studies support that environmental scanning improves an organization's performance (Miller and Friesen 1977, Newgren et al. 1984, Dollinger 1984, West 1988, Daft et al. 1988, Subramanian et al. 1993, Subramanian et al. 1994, Murphy 1987, Ptaszynski 1989). Choo summarizes that ". . . environmental scanning is increasingly being used to drive the strategic planning processes by business and public sector organizations in most developed countries." (Choo 1995, p.101) Additionally, with the development of the Internet, the amount of information easily available to a corporation is vast. "Companies of all shapes and sizes are finding that the Internet provides new opportunities for competitive advantage." (Cronin 1993, p. 40-43)

Information, both external and internal, is a vital part of making strategic decisions (Pawar and Sharda 1997). However, Pawar and Sharda (1997) warn that unsystematic information gathering from the Internet wastes time and money. Hence, one problem becomes acquiring useful information and mining it for what is relevant in a systematic automated manner. Pawar and Sharda (1997) suggest that an Internet-based environmental scanning system can offer benefits, but also has costs. The benefits include the timeliness, low-cost and quantity of the information available. However, the cost of searching can be quite high. Therefore, a second problem is utilizing the information available via the Internet with minimal cost. The environmental scanning

process in this dissertation is developed with the problems discussed in mind. The process of collecting documents can be automated and be as simple as running a few Java programs on a regular basis. As discussed below, we investigate methods to automate the analysis of these documents using information retrieval ideas with discriminant analysis. These methods can also become part of the downloading programs. Thus, the information is gathered and analyzed systematically with minimal cost.

3.1 Problem Setting

As scanning involves gathering and analyzing information from the environment and the Internet provides a vast amount of easily available html documents to scan, it is natural to develop a process to scan html documents for information. There are several approaches that can be employed to solve the problem of analyzing the content of html documents, including traditional data mining or machine learning methods. However, the vector space model (VSM) was developed specifically to solve information retrieval (IR) problems and provides a convenient method for quantifying the problem of content analysis.

Much of the literature relating to the VSM involves improving retrieval results based on a user's query with respect to measures such as precision and recall. However, in this dissertation the model has been adapted to representing web documents in a vector format in order to analyze their content for information about the environment. This content analysis is accomplished by analyzing a set of documents, called a training set. Each document in the training set is classified according to a categorical variable of interest. The information gathered by the training set is then used to determine the value of the categorical variable for future documents. Therefore, the problems that remain are

how to separate documents based on the categorical variable and how to predict the value of the categorical variable for new documents.

As the vector space model is linear, it is natural to separate documents based on a categorical variable with linear methods. One popular linear method is linear discriminant analysis (LDA). LDA is traditionally used for separation of observations into groups and classification of new observations. The various LDA formulations find a linear discriminant function (LDF) based on a training set with known group membership. The LDF can be used to classify new observations. Therefore, LDA solves the problems of how to separate documents based on a categorical variable and how to predict the value of the categorical variable for new documents.

The VSM has been criticized for the mathematical inconsistencies that exist between the various interpretations that have been used in the literature as well as its lack of consistent interpretation and use (Raghavan and Wong 1986). The new adaptation of the model only uses the linear vector representation for documents and terms, without the need for the vector operations that have been criticized for their lack of theoretical justification. Other criticisms are the lack of guidelines for choosing the similarity measure and the assumed orthogonality of terms, an assumption arising from the extreme difficulty of determining term-term correlations (Salton 1989). In our adaptation, no similarity measure is needed, as there are no queries. The literature provides theoretical discussion of how correlated, non-orthogonal term vectors are incorporated into the basic VSM (Raghavan and Wong 1986, Salton 1989), but there are no guidelines on how exactly term-term correlations for correlated, non-orthogonal vectors are determined. In addition, according to Salton "... it is not simple to generate useful term associations."

(Salton 1989, p. 314) In spite of the criticism of the orthogonality of terms, there have been useful, interesting results (Salton 1971, Salton 1975, Salton and McGill 1983). Therefore, we use the simplifying assumption of orthogonality of terms.

In conclusion, the process developed in this dissertation for web-based environmental scanning is simple, directed and automated. The content analysis of the set of collected documents is based on the well-documented, frequently used VSM representation developed for information retrieval combined with traditional LDA for separating the documents based on a categorical variable. Section 3.2 offers a list of research questions based on this process. Section 3.3 describes the application environment used to empirically analyze our environmental scanning process. Finally, Section 3.4 provides an in depth description of the VSM adaptation and methods for determining the linear discriminant function in our application environment.

3.2 Research Questions

Based on the description of the need for an automated web-based environmental scanning process and the explanation of the VSM and LDA as the tools used to develop that process, two general research questions are:

RQ1: How well does the process classify or group the training set of documents based on a categorical variable?

RQ2: Does the process predict the correct classification better than random guessing?

3.3 Application Environment

The relationship between stock returns and news is a hot topic in the financial literature with several very recent papers in the financial literature examining this relationship (Huberman and Regev 2001, Chan in press, Daniel, K., & Titman, S. 2001.

Market reactions to tangible and intangible information. Unpublished manuscript.). Additionally, data about publicly traded companies is readily available via the Internet, so the environmental scanning process developed in this dissertation is empirically tested in the stock market. Online news articles about specific stocks are collected for simple signals, as indicated by the terms used in the articles, about future stock returns or changes in trading volume. A list of web sites used to collect the news articles is provided in Appendix B Table 18. The information collected via the scanning process is the content of the articles. Each article in the training set is analyzed to determine whether the article indicates an increase or a decrease in stock return relative to the market or a change in trading volume as compared to the previous day.

3.4 Vector Space Representation and Discriminant Analysis

For each stock, the k articles collected will be used to determine if the text or terms in the article indicate whether the stock's return will increase or decrease relative to the market in the target period following the report or whether the stock's trading volume will increase or decrease. Once the articles or documents are collected, a set of n index terms needs to be determined. Let $\mathbf{t}_1,\ldots,\mathbf{t}_n$ for $\mathbf{t}_i \in \Re^n$, be the vectors corresponding to the n index terms. The term vectors form a vector space. When the terms are linearly independent, the dimensionality of the vector space is n. Table 4 summarizes the notation for the vector space representation of the problem.

With full dimensionality, each document can be written as a linear combination of term vectors. Articles or documents are represented by vectors, $\mathbf{d}_r$. For the collection of k documents or news articles for each stock, an $n \times k$ document by term matrix $\mathbf{D}$ can be

constructed with the document vectors where each column of the document matrix corresponds to a document vector $\mathbf{d}_r$.

$$\mathbf{d}_r = \sum_{i=1}^{n} a_{i,r}\mathbf{t}_i = \mathbf{T}\mathbf{a}_r$$

$$\mathbf{d}_r = \mathbf{D}\mathbf{e}_r = \mathbf{T}\mathbf{A}\mathbf{e}_r \quad r = 1,\ldots,k$$

The elements of $\mathbf{d}_r$ are the term weights when the terms are orthogonal, which can be confirmed by $\mathbf{t}_i' \mathbf{d}_r = a_{i,r}$. The $a_{i,r}$'s are determined by an indexing operation on the document collection.

Table 4. Variable definition

Variable	Meaning
k	Number of documents
k_i	Number of documents in group i, $i = 1,2$, $k = k_1 + k_2$
n	Number of terms
$\mathbf{t}_i$	$n \times 1$ term vector representing term i
$\mathbf{T}$	$n \times n$ term matrix where $\mathbf{t}_i$'s are the columns
$\mathbf{d}_r$	$n \times 1$ vector for document r, $\mathbf{d}_r = \mathbf{D}\mathbf{e}_r$
$\mathbf{D}$	$n \times k$ matrix where $\mathbf{d}_r$'s are the columns
$\mathbf{D}_i$	$n \times k_i$ matrix where $\mathbf{d}_r$'s in group i are the columns, $i = 1,2$
$\mathbf{A}$	$n \times k$ term document matrix where $a_{i,r}$ is the weight of term i in document r
$\mathbf{a}_r$	$n \times 1$ term weight vector for document r, $\mathbf{a}_r = \mathbf{A}\mathbf{e}_r$
$\mathbf{G}_t$	$n \times n$ term-term correlation matrix where $g_{i,j} = \mathbf{t}_i \cdot \mathbf{t}_j$ is the correlation between term $\mathbf{t}_i$ and $\mathbf{t}_j$
$\mathbf{G}_d$	$k \times k$ document-document correlation matrix where $g_{r,s} = \mathbf{d}_r \cdot \mathbf{d}_s$ is the correlation between document $\mathbf{d}_r$ and $\mathbf{d}_s$
π_i	Documents in group i, $i = 1,2$
z	Cutting score for Fisher discriminant analysis
q	Distance variable in for LP formulation of discriminant analysis
$\mathbf{q}_i$	$k_i \times 1$ vector of distances for documents in group $i = 1,2$
$\mathbf{1}_i$	Column vector $N_i \times 1$ of ones, $i = 1, 2$

Once the documents have the above representation for each stock, each document needs to be classified according one of two methods. The first method classifies documents as first appearing a day prior to the stock return increasing relative to the market or as first appearing a day prior to the stock return decreasing relative to the market. All the documents that correspond to an increase or no change in the stock's return are group 1 documents, π_1, and the documents that correspond to a decrease in return are group 2 documents, π_2. The second method classifies documents as first appearing a day prior to the stock's trading volume increasing as compared to the previous day or as first appearing a day prior to the stock's trading volume decreasing. All the documents that correspond to an increase in the stock's trading volume are group 1 documents, π_1, and the documents that correspond to a decrease or no change in volume are group 2 documents, π_2. Hence, all documents are classified as $\mathbf{d}_i$, $i = 1$ if $\mathbf{d}_r \in \pi_1$ and $i = 2$ if $\mathbf{d}_r \in \pi_2$. Let k_1 be the number of group 1 documents so that $k_2 = k - k_1$ is the number of group 2 documents. Let $\mathbf{d}_1, \ldots, \mathbf{d}_{k_1}$ be the group 1 documents and $\mathbf{d}_{k_1+1}, \ldots, \mathbf{d}_k$ be the group 2 documents.

Discriminant analysis is then used to derive a variate, $\mathbf{w}'\mathbf{d}_r$, that best discriminates between the groups. In the classic Fisher approach, the elements of $\mathbf{w}$ are weights that are determined by maximizing the between group variance relative to the within group variance. The discriminant score, z_r, is calculated for each document by $z_r = \mathbf{w}'\mathbf{d}_r$. The score is used to predict whether the document is in group 1 or group 2 according to the cutting score z by

$$\begin{aligned} \mathbf{w}'\mathbf{d}_r &\geq z \quad \mathbf{d}_r \in \pi_1 \\ \mathbf{w}'\mathbf{d}_r &\leq z \quad \mathbf{d}_r \in \pi_2 \end{aligned}. \tag{1}$$

The cutting score, z, and the weights, $\mathbf{w}$, are determined according to the methods described in Section 2.3 on discriminant analysis.

When terms are uncorrelated and orthogonal

$$\mathbf{t}_i' \mathbf{t}_j = \delta(i = j).$$

Hence,

$$\mathbf{T} = \mathbf{I}.$$

The equations in (1) simplify to

$$\begin{aligned} \mathbf{w}'\mathbf{a}_r &\geq z \quad \mathbf{d}_r \in \pi_1 \\ \mathbf{w}'\mathbf{a}_r &\leq z \quad \mathbf{d}_r \in \pi_2 \end{aligned}.$$

Fisher's sample linear discriminant function (Johnson and Wichern 1982) is given by

$$y = \left(\bar{\mathbf{d}}_1 - \bar{\mathbf{d}}_2\right)' \mathbf{S}_{\text{pooled}}^{-1} \mathbf{d}_r \tag{2}$$

where

$$\mathbf{S}_{\text{pooled}} = \frac{(k_1 - 1)\mathbf{S}_1^2 + (k_2 - 1)\mathbf{S}_2^2}{(k - 2)},$$

$$\bar{\mathbf{d}}_1 = \frac{1}{k_1} \sum_{j=1}^{k_1} \mathbf{d}_j,$$

$$\bar{\mathbf{d}}_2 = \frac{1}{k - k_1} \sum_{j=k_1}^{k} \mathbf{d}_j,$$

$$\mathbf{S}_1^2 = \frac{1}{k_1 - 1} \sum_{j=1}^{k_1} \left(\mathbf{d}_1 - \bar{\mathbf{d}}_1\right)\left(\mathbf{d}_1 - \bar{\mathbf{d}}_1\right)'$$

and

$$\mathbf{S}_2^2 = \frac{1}{k_2 - 1} \sum_{j=k_1}^{k} \left(\mathbf{d}_2 - \bar{\mathbf{d}}_2\right)\left(\mathbf{d}_2 - \bar{\mathbf{d}}_2\right)'$$

which simplifies to

$$y = \left(\bar{\mathbf{a}}_1 - \bar{\mathbf{a}}_2\right)' \mathbf{S}_{\text{pooled}}^{-1} \mathbf{a}_r$$

where the cutting score is given by

$$z = \frac{1}{2}\left(\bar{\mathbf{d}}_1 - \bar{\mathbf{d}}_2\right)' \mathbf{s}_{\text{pooled}}^{-1} \left(\bar{\mathbf{d}}_1 + \bar{\mathbf{d}}_2\right). \tag{3}$$

CHAPTER 4
ENVIRONMENTAL SCANNING PROCESS

The purpose of this chapter is to provide further details of the methodology developed in this dissertation. The purpose of this dissertation is to automate the process of environmental scanning on the World Wide Web via the use of web-based methods. This will be done in a series of steps described in some detail in this chapter. A summary of these steps is provided in Figure 1. A detailed discussion of each step is given in Section 4.1. Section 4.2 outlines the plan for testing the data collected.

4.1 Summary

The first step involves automating the collection of web documents. A computer program in Java was written that automatically collects potentially relevant documents from previously identified web sites given in Appendix B Table 18. The documents are news articles pertaining to a predefined set of stocks, given in Appendix B Table 19. The program visits each of these sites on a daily basis and collects news articles related to specific stocks on the list of stocks. The html documents are stored locally in files. Each document for each stock can be classified into one of two groups using the current stock return or trading volume as the classification mechanism. By the first classification mechanism, news documents appearing before stock returns increase (decrease) relative to the market return are classified as π_1 (π_2), where the market return is measured by the S&P500 Index. Alternatively, news documents appearing before trading volume of a stock increases (decreases) are classified as π_1 (π_2). The intuition is the contents of the

news documents signal an increase or decrease in stock returns relative to the market or signal a change in trading volume.

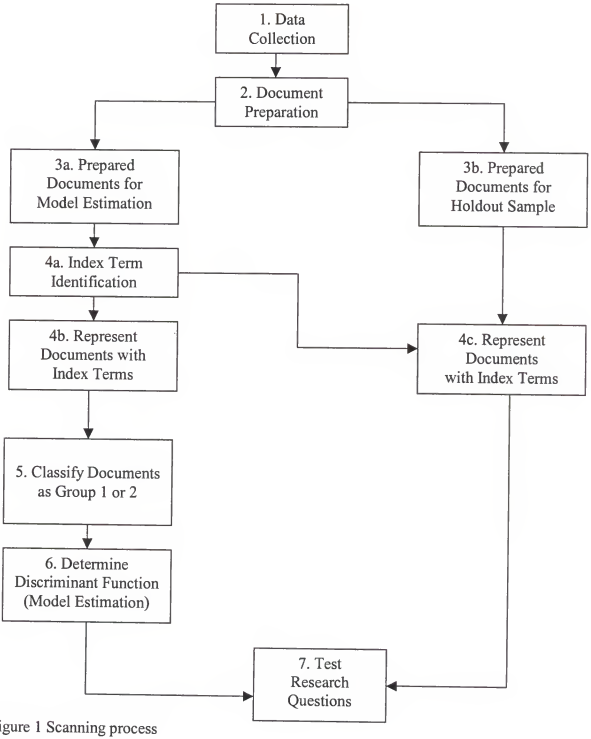

Figure 1 Scanning process

The second step in the procedure involves cleaning the html documents. Cleaning the documents involves removing html tags and words in a stopword list (Frakes and Baeza-Yates 1992). A stopword list consists of words that are very frequent in the English language and do not have discriminatory meaning, such as the words "the" and "of". The list of stopwords is given in Appendix B Table 20. As the words in the stopword list are very frequent in document collections, they do not help distinguish documents from one another and may be removed from the document collection. Cleaning the documents involves running a Java program. Additionally, documents are stemmed, the process of replacing words with their word stems, using the Porter stemming algorithm (Porter 1980).

Once the documents have been cleaned and stemmed, they are indexed. "The process of constructing document surrogates by assigning identifiers to text items is known as indexing." (Salton 1989, p. 275) The idea behind indexing is to find the set of index terms that best represent the documents in the collection. The method of indexing chosen determines how the documents will be represented in vector format. Indexing is done in accordance with the algorithm discussed in Section 2.2 using term weights computed by multiplying term frequency by the inverse document frequency. Indexing is accomplished via running a series of three Java programs. The first creates a master list of terms in the entire document collection for each stock. The second computes the inverse document frequency of each term in the document collection, only including terms that occur five or more times in the entire document collection. Finally, each document is represented numerically by counting the term frequency for each distinct

stem in the document and multiplying by the corresponding inverse document frequency for that term.

After the documents have been indexed, they are in vector format. Next, their classification, π_1 or π_2, will be determined. The first classification method involves measuring stock returns relative to the market. Let r_i be the rate of return of stock i and r_m be the rate of return of the market as measured by the S&P500 Index. A stock's performance relative to the market is measured by computing the difference between the stock's rate of return r_i and the market's rate of return r_m. News documents appearing a day before $r_i - r_m \geq 0$ are classified as π_1 and documents appearing a day before $r_i - r_m < 0$ are classified as π_2.

The second method of classification is based on trading volume. News documents appearing a day before an increase in a stock's trading volume as compared to the previous day's trading volume are classified as π_1 and documents appearing a day before a decrease or no change in trading volume are classified as π_2.

After classification, each document is in a form that can be analyzed via discriminant analysis. The set of independent variables for the discriminant procedure is given by the set of index terms and the dependent variable is either classification based on return or trading volume. Each document is an observation in the discriminant procedure. The vector representation and discriminant analysis were discussed in Section 3.4. Discriminant analysis is performed in two steps. First, a forward stepwise discriminant procedure is employed with the significance level, set at 15%, of an F test from an analysis of covariance used as the selection criteria for a variable to enter the model. Next, Fisher's (1936) linear discriminant model is determined by calculating the

linear discriminant coefficients, $\mathbf{w}$, using the variables that entered the model via the stepwise procedure and calculating the cutting score, z.

Once the discriminant function is determined, the validity of the discriminant model will be checked via a holdout sample. Documents in the holdout sample will be used to predict whether stock returns will increase or decrease relative to the market or whether trading volume will increase or decrease. The values of z and $\mathbf{w}$ computed in the discriminant procedure will be updated periodically according to the newly collected data. The plan for testing the process discussed in this section is given in Section 4.2.

4.2 Experimental Plan

The document collection is divided into an 80% training sample and a 20% holdout sample by using the first 80% of the document collected in time in the training sample and the next 20% in the holdout sample. Classification matrices for the training and the holdout samples are examined. To assess external validity of the model statistical significance for the holdout sample classification matrix is determined via Press's Q using a chi-square distribution with one degree of freedom for two-group classification. Press's Q is given by

$$\text{Press's Q} = \frac{(N - nK)^2}{N(K-1)}$$

where N is the total sample size, n is the number of correctly classified observations and K is the number of groups (Hair et al. 1998).

The internal validity of the linear discriminant model is checked by examining the classification matrix for the training set and by examining the classification matrix determined via a jackknife cross-validation procedure. The proportional chance criterion (Hair et al. 1998) is calculated for both classification matrices. The proportional chance

criterion is $p_c = p^2 + (1-p)^2$ where p is the proportion of group 1 documents in the training set. The proportional chance criterion is compared to the hit ratio, p_h, defined as the proportion of documents that are classified correctly by the discriminant model. If the hit ratio is statistically significantly larger than the proportional chance criterion according to the z-statistic

$$z = \frac{(p_h - p_c)}{\sqrt{\frac{p_c(1-p_c)}{N}}}$$

then the classification by the jackknife cross-validation is statistically better than chance.

Testing the training set and the cross-validation classification matrices via the proportional chance criterion addresses research question one (RQ1) given in Section 3.2. Testing the holdout sample addresses research question two (RQ2).

Table 5. Summary of experimental plan

1. Check 80% training sample for validity via classification matrix and leave-one-out classification
2. Check 20% holdout sample for validity via classification matrix
3. Compare the performance of the two classification mechanisms
4. Check 20% holdout sample results for improvement with independent variables augmented with x_1

In addition, as a benchmark, classification of the documents is done randomly, as opposed to using actual daily changes in stock returns or trading volume for determining group membership. Performance of the model on random classification is compared to performance using actual classification mechanisms to determine if the actual classification performs statistically better than random classification.

Finally, an independent variable, x_1, is added to the set of independent variables. The variable represents the stock's performance in the three days prior to the date used

for classification. The variable, x_1, is calculated by analyzing the classification of the stock based on return or trading volume over these three days. Table 6 summarizes the values of x_1. Group membership is assessed for these three days, resulting in eight possible combinations. To compute the value of x_1, group membership is coded in binary, group one corresponds to a one and group two to a zero. The value of x_1 is given by

$$x_1 = x_1^1(2^2) + x_1^2(2^1) + x_1^3(2^0).$$

An advantage of using this method of calculating x_1 is that more weight is given to the most recent day's classification.

Table 6. Summary of values of x_1

One Day Prior x_1^1	Two Days Prior x_1^2	Three Days Prior x_1^3	Value of x_1
1	1	1	7
1	1	2	6
1	2	1	5
1	2	2	4
2	1	1	3
2	1	2	2
2	2	1	1
2	2	2	0

In light of the experimental plan, two additional research questions are:

RQ3: Which method of classification works best, the stock return method or the method based on trading volume changes?

RQ4: Do holdout sample results improve when adding x_1 as an independent variable?

A summary of the experimental plan is given in Table 5. Each numbered item in Table 5 addresses the same numbered research question. The results of the plan are given in Chapter 5.

CHAPTER 5
RESULTS

The purpose of this chapter is to discuss the results of the data analysis outlined in Section 4.2. There were 186 stocks in the original data set listed in Appendix B Table 19. Stocks with 200 articles or less were not considered as they did not even average more than one article per day. This provides a sample of 96 stocks remaining in the analysis. Additionally, stocks that did not trade on all days that data was collected were removed from the list of stocks for analysis, leaving 93 stocks.

The chapter starts with an overview of the document collections for each stock in Section 5.1. Section 5.2 provides results for the 80% training set. Section 5.3 provides results for the 20% holdout sample. Section 5.4 provides results for the case where the set of independent variables is augmented with a variable representing the stock's performance in the previous three days.

5.1 Summary of Document Collections

Table 7 provides a summary of the size of the training set (training no. of docs), the size of the holdout set of documents (holdout no. of docs) and the number of word stems remaining after indexing (no. of stems). The average training set size is 974, the average holdout sample size is 195 and the average number of word stems is 5217. Table 7 also provides the number of group 1 documents in the training set (training group 1), the number of group 2 documents in the training set (training group 2), the number of

group 1 documents in the holdout sample (holdout group 1) and the number of group 2 documents in the holdout sample (holdout group 2).

Recall that the maximum number of entering variables in the step-wise discriminant procedure is the total number of documents in the collection for a given stock divided by four. When using classification based on return, 39 of the stocks entered the maximum number of variables allowed. When using classification based on volume, 41 of the stocks entered the maximum number of variables allowed.

5.2 Training Set Results

Table 8 provides the classification matrix for the training set of documents for the stocks with classification based in stock returns. The average hit ratio, defined as the percent of correctly classified documents in the collection, is 92.11% for Table 8. Using the z-statistic based on the proportional chance criterion defined in Section 4.2, every stock has a classification matrix with correct classification that is significant at 1%.

Table 9 provides the jackknife cross-validation classification matrix for the training set of documents for stocks with classification based on stock returns. The average hit ratio is 89.06% for Table 9. The z-statistic for the jackknife cross-validation classification matrix for every stock is significant at 1%.

Table 10 provides the classification matrix for the training set of documents for the stocks with classification based on changes in volume. The average hit ratio is 91.68% for Table 10. Using the proportional chance criterion, every stock has a classification matrix with correct classification that is significant at 1%.

Table 11 provides the jackknife cross-validation classification matrix for the training set of documents for stocks with classification based on changes in volume. The

average hit ratio is 88.68% for Table 11. Using the proportional chance criterion, every stock has a leave-one-out cross-validation classification matrix with correct classification that is significant at 1%.

5.3 Holdout Sample Results

Table 12 provides a summary of the results of classification based on return for the 20% holdout sample. The average hit ratio is 50.25% for Table 12. Using Press's Q, 16 out of 93 stocks have a holdout classification matrix with correct classification that is significant at 10%. As a benchmark, a subset of 71 of the documents was classified randomly, with the classification proportional to the number of documents in each group. With random classification, only 5 stocks out of 71 had a holdout classification matrix that has significant correct classification at 10%. Originally, 13 out of the 71 stocks had a holdout classification matrix with correct classification significant at 10%. The p-value for the difference between the two proportions, 13 out of 71 and 5 out of 71, for a one-tailed test is 0.0217.

Table 13 provides a summary of the results of classification based on volume for the 20% holdout sample. The average hit ratio is 49.67% for Table 13. Using Press's Q, 16 out of 93 stocks have a holdout classification matrix with correct classification that is significant at 10%. As a benchmark, a subset of the documents was classified randomly, with the classification proportional to the number of documents in each group. With random classification, only 5 stocks out of 72 had a holdout classification matrix that has significant correct classification at 10%. Originally, 13 out of the 72 stocks had a holdout classification matrix with correct classification significant at 10%. The p-value for the

difference between the two proportions, 13 out of 72 and 5 out of 72, for a one-tailed test is 0.0217.

5.4 Prior Classification Variable Results

Table 14 provides the classification matrix for the training set of documents for the stocks with classification based on stock returns with the set of independent variables augmented with x_1, a variable representing the classification of the stock for the three days prior to the date used for classification of the dependent variable. Only stocks that had x_1 enter the model during the step-discriminant procedure, a collection of 82 stocks, are included in Table 14. The average hit ratio, defined as the percent of correctly classified documents in the collection, is 93.09% for Table 14. Using the z-statistic based on the proportional chance criterion defined in Section 4.2, every stock has a classification matrix with correct classification that is significant at 1%. Additionally, the average step number that x_1 entered the model in the step-discriminant procedure is 41.48.

Table 15 provides the classification matrix for the training set of documents for the stocks with classification based on changes in trading volume with the set of independent variables augmented with x_1. Only stocks that had x_1 enter the model during the step-discriminant procedure, a collection of 91 stocks, are included in Table 15. The average hit ratio, defined as the percent of correctly classified documents in the collection, is 94.28% for Table 15. Using the z-statistic based on the proportional chance criterion defined in Section 4.2, every stock has a classification matrix with correct classification that is significant at 1%. Additionally, the average step number that x_1 entered the model in the step-discriminant procedure is 4.12.

Table 16 provides a summary of the results of classification based on return with the set of independent variables augmented with x_1 for the 20% holdout sample. The average hit ratio is 49.91% for Table 16. Using Press's Q, 15 out of 82 stocks have a holdout classification matrix with correct classification that is significant at 10%.

Table 17 provides a summary of the results of classification based on volume with the set of independent variables augmented with x_1 for the 20% holdout sample. The average hit ratio is 52.92% for Table 17. Using Press's Q, 32 out of 91 stocks have a holdout classification matrix with correct classification that is significant at 10%.

Table 7. Return classification collection summary

Stock	No. of Stems	Training No. of Docs	Training Group 1	Training Group 2	Holdout No. of Docs	Holdout Group 1	Holdout Group 2
AAPL	7887	1356	671	685	221	151	70
ABS	3870	548	288	260	135	75	60
ADSX	2133	238	106	132	114	30	84
AEOS	3179	422	227	195	73	52	21
AET	4658	724	426	298	161	75	86
AFFX	2379	245	142	103	29	21	8
ANF	3795	474	305	169	60	41	19
AOL	10347	2623	1156	1467	570	308	262
ASKJ	3089	329	144	185	59	37	22
AXP	8326	1543	884	659	368	168	200
BA	8515	2557	1344	1213	577	312	265
BAB	4506	1028	495	533	233	87	146
BC	2013	214	127	87	56	12	44
BDK	2712	296	160	136	87	30	57
BMY	7786	1762	869	893	444	238	206
BUD	4312	709	336	373	114	72	42
C	8832	2257	1209	1048	496	241	255
CCU	4357	594	293	301	109	57	52
CHIR	3526	447	241	206	52	34	18
CMTN	3538	220	110	110	25	14	11
COX	6292	1097	565	532	114	77	37
CREE	2074	226	89	137	98	41	57
CSCO	10038	2171	1102	1069	389	167	222
CVC	5254	704	354	350	86	33	53
DAL	7540	2158	1122	1036	427	207	220
DD	7301	1122	633	489	232	135	97
DELL	7920	1717	903	814	312	185	127
DOW	5129	849	445	404	96	72	24
DRI	2460	287	199	88	77	42	35
EBAY	7490	1298	618	680	213	131	82
ELY	3686	472	234	238	81	52	29
ERICY	6704	1330	769	561	290	160	130
F	9589	2795	1314	1481	543	278	265
FDX	6215	1055	625	430	245	110	135
FPL	1601	253	133	120	110	73	37
GD	5509	1149	581	568	251	126	125
GE	10249	2413	1100	1313	520	305	215
GLW	7275	1208	516	692	231	61	170
GM	10002	2744	1384	1360	596	352	244
GPS	7483	1155	446	709	158	101	57
GR	3549	539	252	287	80	28	52

Table 7. Continued

Stock	No. of Stems	Training No. of Docs	Training Group 1	Training Group 2	Holdout No. of Docs	Holdout Group 1	Holdout Group 2
GTW	6540	1051	463	588	182	91	91
HDI	3233	333	156	177	57	49	8
HUM	3115	407	249	158	36	14	22
IBM	10441	2426	1028	1398	346	209	137
ILXO	2082	206	93	113	17	4	13
INTC	9728	2359	1311	1048	527	207	320
JNJ	6992	1370	693	677	344	176	168
KM	8200	1628	795	833	472	258	214
KR	4399	559	294	265	159	87	72
LIZ	2368	252	147	105	64	19	45
LLTC	2860	299	133	166	43	37	6
LLY	6702	1396	766	630	244	124	120
LMT	6700	1550	791	759	399	159	240
LUV	6133	1169	672	497	209	94	115
MCD	7148	1406	720	686	278	135	143
MO	7930	1514	831	683	399	203	196
MON	3983	610	339	271	190	94	96
MOT	9285	2090	949	1141	379	171	208
MRK	8083	1282	673	609	377	180	197
MYG	3081	809	413	396	85	45	40
NKE	3271	415	212	203	89	21	68
NOK	5812	1053	548	505	210	124	86
NSANY	2166	432	240	192	140	49	91
NVS	4160	722	366	356	182	106	76
NXTL	5195	831	425	406	146	89	57
ORCL	6098	1183	597	586	226	174	52
PBG	2104	315	179	136	63	47	16
PD	1934	294	198	96	67	24	43
PEP	5226	932	382	550	157	88	69
PHA	4545	737	393	344	205	121	84
RBK	2532	321	144	177	59	37	22
RL	2073	204	117	87	19	13	6
ROST	1899	216	84	132	17	9	8
S	6390	1008	661	347	153	37	116
SNE	5762	1006	563	443	199	95	104
SO	2238	273	104	169	90	36	54
SOI	2546	269	176	93	28	15	13
TGT	6759	1011	624	387	138	56	82
TM	3673	732	365	367	163	69	94

Table 7. Continued

Stock	No. of Stems	Training No. of Docs	Training Group 1	Training Group 2	Holdout No. of Docs	Holdout Group 1	Holdout Group 2
TMPW	4366	670	323	347	60	33	27
TOM	2330	225	101	124	16	5	11
TXN	5635	977	515	462	226	135	91
TXU	3033	480	202	278	60	47	13
USAI	4579	707	325	382	108	30	78
WEN	2984	428	165	263	90	30	60
WHR	3064	424	229	195	86	45	41
WIN	2044	228	130	98	68	27	41
WMT	8023	1670	904	766	319	132	187
WPPGY	2624	349	185	164	69	24	45
X	3158	549	357	192	51	29	22
XEL	2290	277	133	144	131	100	31
XOM	7610	1473	722	751	277	121	156

Table 8. Return classification training set results

Stock	Hit Ratio	Chance	Z-Statistic	Significant
AAPL	90.93%	0.50	30.14	yes
ABS	94.53%	0.50	20.79	yes
ADSX	88.66%	0.51	11.74	yes
AEOS	86.26%	0.50	14.78	yes
AET	92.96%	0.52	22.29	yes
AFFX	95.92%	0.51	13.98	yes
ANF	89.66%	0.54	15.53	yes
AOL	97.18%	0.51	47.61	yes
ASKJ	90.88%	0.51	14.55	yes
AXP	94.10%	0.51	33.82	yes
BA	97.11%	0.50	47.51	yes
BAB	96.30%	0.50	29.65	yes
BC	91.59%	0.52	11.66	yes
BDK	85.47%	0.50	12.09	yes
BMY	94.89%	0.50	37.68	yes
BUD	83.36%	0.50	17.69	yes
C	95.30%	0.50	42.80	yes
CCU	90.91%	0.50	19.94	yes
CHIR	93.29%	0.50	18.18	yes
CMTN	87.73%	0.50	11.19	yes
COX	89.61%	0.50	26.21	yes
CREE	95.58%	0.52	13.04	yes
CSCO	95.72%	0.50	42.59	yes
CVC	88.49%	0.50	20.43	yes
DAL	95.92%	0.50	42.59	yes
DD	90.37%	0.51	26.50	yes
DELL	93.83%	0.50	36.21	yes
DOW	91.99%	0.50	24.40	yes
DRI	90.24%	0.57	11.23	yes
EBAY	90.91%	0.50	29.40	yes
ELY	92.58%	0.50	18.50	yes
ERICY	92.93%	0.51	30.43	yes
F	94.06%	0.50	46.40	yes
FDX	91.37%	0.52	25.78	yes
FPL	96.84%	0.50	14.86	yes
GD	92.08%	0.50	28.52	yes
GE	93.37%	0.50	42.23	yes
GLW	93.96%	0.51	29.82	yes
GM	95.41%	0.50	47.57	yes
GPS	90.82%	0.53	26.02	yes
GR	89.05%	0.50	18.04	yes
GTW	92.29%	0.51	26.97	yes
HDI	85.89%	0.50	13.02	yes

Table 8. Continued

Stock	Hit Ratio	Chance	Z-Statistic	Significant
HUM	91.40%	0.52	15.72	yes
IBM	94.77%	0.51	42.96	yes
ILXO	92.23%	0.50	11.99	yes
INTC	95.13%	0.51	43.23	yes
JNJ	94.16%	0.50	32.69	yes
KM	92.51%	0.50	34.28	yes
KR	82.47%	0.50	15.29	yes
LIZ	90.08%	0.51	12.29	yes
LLTC	85.95%	0.51	12.22	yes
LLY	94.48%	0.50	32.89	yes
LMT	95.87%	0.50	36.10	yes
LUV	91.62%	0.51	27.70	yes
MCD	92.18%	0.50	31.61	yes
MO	93.92%	0.50	33.81	yes
MON	92.30%	0.51	20.59	yes
MOT	94.40%	0.50	40.21	yes
MRK	94.31%	0.50	31.64	yes
MYG	85.78%	0.50	20.34	yes
NKE	89.16%	0.50	15.94	yes
NOK	95.06%	0.50	29.19	yes
NSANY	95.14%	0.51	18.51	yes
NVS	96.95%	0.50	25.23	yes
NXTL	91.46%	0.50	23.89	yes
ORCL	96.11%	0.50	31.72	yes
PBG	83.17%	0.51	11.45	yes
PD	97.62%	0.56	14.37	yes
PEP	89.27%	0.52	23.00	yes
PHA	94.84%	0.50	24.23	yes
RBK	84.42%	0.51	12.15	yes
RL	85.78%	0.51	9.92	yes
ROST	81.02%	0.52	8.40	yes
S	91.96%	0.55	23.68	yes
SNE	96.32%	0.51	28.94	yes
SO	95.97%	0.53	14.28	yes
SOI	96.28%	0.55	13.68	yes
TGT	91.49%	0.53	24.68	yes
TM	96.72%	0.50	25.28	yes
TMPW	93.43%	0.50	22.45	yes
TOM	90.22%	0.51	11.91	yes
TXN	95.80%	0.50	28.54	yes
TXU	96.04%	0.51	19.63	yes
USAI	92.64%	0.50	22.51	yes

Table 8. Continued

Stock	Hit Ratio	Chance	Z-Statistic	Significant
WEN	86.68%	0.53	14.11	yes
WHR	92.69%	0.50	17.45	yes
WIN	88.60%	0.51	11.36	yes
WMT	95.57%	0.50	36.97	yes
WPPGY	97.99%	0.50	17.86	yes
X	89.25%	0.55	16.34	yes
XEL	90.25%	0.50	13.37	yes
XOM	94.37%	0.50	34.04	yes

Table 9. Return classification training set leave-one-out cross-validation

Stock	Hit Ratio	Chance	Z-Statistic	Significant
AAPL	87.39%	0.50	27.53	yes
ABS	90.51%	0.50	18.91	yes
ADSX	82.77%	0.51	9.93	yes
AEOS	83.89%	0.50	13.80	yes
AET	90.88%	0.52	21.17	yes
AFFX	93.47%	0.51	13.22	yes
ANF	86.50%	0.54	14.15	yes
AOL	94.97%	0.51	45.35	yes
ASKJ	86.32%	0.51	12.90	yes
AXP	91.12%	0.51	31.48	yes
BA	94.80%	0.50	45.17	yes
BAB	93.58%	0.50	27.90	yes
BC	89.72%	0.52	11.12	yes
BDK	77.03%	0.50	9.19	yes
BMY	92.96%	0.50	36.06	yes
BUD	81.10%	0.50	16.49	yes
C	92.25%	0.50	39.90	yes
CCU	89.06%	0.50	19.03	yes
CHIR	89.71%	0.50	16.66	yes
CMTN	85.45%	0.50	10.52	yes
COX	87.51%	0.50	24.82	yes
CREE	92.48%	0.52	12.11	yes
CSCO	93.74%	0.50	40.75	yes
CVC	84.52%	0.50	18.32	yes
DAL	92.96%	0.50	39.84	yes
DD	87.52%	0.51	24.59	yes
DELL	90.74%	0.50	33.65	yes
DOW	89.52%	0.50	22.96	yes
DRI	87.46%	0.57	10.27	yes
EBAY	87.83%	0.50	27.17	yes
ELY	88.56%	0.50	16.75	yes
ERICY	89.32%	0.51	27.80	yes
F	91.20%	0.50	43.37	yes
FDX	88.82%	0.52	24.12	yes
FPL	92.09%	0.50	13.35	yes
GD	89.12%	0.50	26.52	yes
GE	90.14%	0.50	39.05	yes
GLW	91.06%	0.51	27.81	yes
GM	93.00%	0.50	45.05	yes
GPS	87.97%	0.53	24.08	yes
GR	87.57%	0.50	17.35	yes
GTW	90.10%	0.51	25.55	yes
HDI	83.18%	0.50	12.04	yes

Table 9. Continued

Stock	Hit Ratio	Chance	Z-Statistic	Significant
HUM	87.71%	0.52	14.23	yes
IBM	91.59%	0.51	39.84	yes
ILXO	86.41%	0.50	10.32	yes
INTC	92.20%	0.51	40.39	yes
JNJ	90.80%	0.50	30.20	yes
KM	89.19%	0.50	31.60	yes
KR	78.18%	0.50	13.26	yes
LIZ	84.92%	0.51	10.65	yes
LLTC	78.26%	0.51	9.56	yes
LLY	92.62%	0.50	31.50	yes
LMT	94.32%	0.50	34.88	yes
LUV	90.16%	0.51	26.70	yes
MCD	88.55%	0.50	28.89	yes
MO	90.82%	0.50	31.40	yes
MON	88.85%	0.51	18.89	yes
MOT	90.91%	0.50	37.02	yes
MRK	92.75%	0.50	30.52	yes
MYG	77.00%	0.50	11.13	yes
NKE	84.82%	0.50	14.18	yes
NOK	93.83%	0.50	28.39	yes
NSANY	93.75%	0.51	17.93	yes
NVS	95.43%	0.50	24.41	yes
NXTL	89.05%	0.50	22.50	yes
ORCL	94.67%	0.50	30.73	yes
PBG	82.22%	0.51	11.11	yes
PD	96.60%	0.56	14.02	yes
PEP	87.02%	0.52	21.62	yes
PHA	92.94%	0.50	23.20	yes
RBK	79.75%	0.51	10.47	yes
RL	80.88%	0.51	8.51	yes
ROST	77.31%	0.52	7.31	yes
S	87.40%	0.55	20.77	yes
SNE	94.63%	0.51	27.86	yes
SO	91.94%	0.53	12.94	yes
SOI	92.19%	0.55	12.34	yes
TGT	87.44%	0.53	22.09	yes
TM	95.22%	0.50	24.47	yes
TMPW	90.15%	0.50	20.75	yes
TOM	87.56%	0.51	11.11	yes
TXN	93.76%	0.50	27.26	yes
TXU	94.38%	0.51	18.90	yes
USAI	89.25%	0.50	20.70	yes

Table 9. Continued

Stock	Hit Ratio	Chance	Z-Statistic	Significant
WEN	81.31%	0.53	11.89	yes
WHR	89.39%	0.50	16.09	yes
WIN	86.40%	0.51	10.70	yes
WMT	93.35%	0.50	35.15	yes
WPPGY	96.28%	0.50	17.22	yes
X	86.89%	0.55	15.23	yes
XEL	88.81%	0.50	12.89	yes
XOM	92.80%	0.50	32.84	yes

Table 10. Volume classification training set results

Stock	Hit Ratio	Chance	Z-Statistic	Significant
AAPL	89.53%	0.50	28.79	yes
ABS	94.11%	0.50	20.49	yes
ADSX	90.21%	0.56	10.66	yes
AEOS	83.85%	0.50	13.86	yes
AET	93.14%	0.51	22.65	yes
AFFX	94.21%	0.51	13.55	yes
ANF	88.98%	0.51	16.54	yes
AOL	96.49%	0.50	47.05	yes
ASKJ	86.89%	0.50	13.35	yes
AXP	94.12%	0.52	32.59	yes
BA	96.83%	0.50	46.71	yes
BAB	96.75%	0.50	29.73	yes
BC	89.15%	0.51	11.22	yes
BDK	84.12%	0.51	11.52	yes
BMY	93.87%	0.50	36.26	yes
BUD	86.23%	0.50	19.11	yes
C	94.45%	0.50	41.20	yes
CCU	93.32%	0.50	20.92	yes
CHIR	94.83%	0.51	18.62	yes
CMTN	80.00%	0.50	8.80	yes
COX	89.74%	0.50	26.12	yes
CREE	96.38%	0.50	13.78	yes
CSCO	95.29%	0.50	41.71	yes
CVC	88.24%	0.51	19.82	yes
DAL	96.07%	0.51	41.19	yes
DD	89.50%	0.50	26.35	yes
DELL	94.36%	0.50	36.04	yes
DOW	90.28%	0.51	22.48	yes
DRI	86.06%	0.53	11.09	yes
EBAY	90.81%	0.50	29.25	yes
ELY	87.15%	0.51	15.84	yes
ERICY	93.06%	0.50	31.18	yes
F	94.81%	0.50	46.84	yes
FDX	91.15%	0.50	26.42	yes
FPL	95.34%	0.51	13.77	yes
GD	94.23%	0.52	28.81	yes
GE	94.96%	0.50	43.27	yes
GLW	92.64%	0.51	28.83	yes
GM	95.46%	0.51	46.34	yes
GPS	87.48%	0.50	25.23	yes
GR	89.89%	0.50	18.43	yes
GTW	91.38%	0.51	25.79	yes
HDI	89.06%	0.51	13.89	yes

Table 10. Continued

Stock	Hit Ratio	Chance	Z-Statistic	Significant
HUM	90.64%	0.50	16.37	yes
IBM	95.64%	0.50	43.50	yes
ILXO	91.22%	0.51	11.65	yes
INTC	94.38%	0.50	42.59	yes
JNJ	93.72%	0.51	31.65	yes
KM	93.80%	0.50	34.79	yes
KR	82.82%	0.52	14.29	yes
LIZ	89.26%	0.50	12.21	yes
LLTC	88.89%	0.50	13.40	yes
LLY	92.74%	0.50	31.69	yes
LMT	95.16%	0.51	34.69	yes
LUV	92.33%	0.50	28.66	yes
MCD	93.61%	0.50	31.86	yes
MO	92.09%	0.51	31.52	yes
MON	92.42%	0.50	20.80	yes
MOT	94.90%	0.50	40.70	yes
MRK	94.84%	0.52	30.55	yes
MYG	87.36%	0.52	20.08	yes
NKE	85.71%	0.50	14.51	yes
NOK	94.97%	0.51	28.46	yes
NSANY	97.18%	0.50	19.44	yes
NVS	96.39%	0.50	24.84	yes
NXTL	91.39%	0.52	22.36	yes
ORCL	97.35%	0.50	32.35	yes
PBG	82.37%	0.50	11.33	yes
PD	97.61%	0.55	14.58	yes
PEP	89.88%	0.50	24.18	yes
PHA	93.99%	0.51	23.18	yes
RBK	84.03%	0.50	12.04	yes
RL	83.92%	0.52	9.09	yes
ROST	80.37%	0.54	7.76	yes
S	90.61%	0.50	25.69	yes
SNE	96.79%	0.50	29.47	yes
SO	93.75%	0.50	14.43	yes
SOI	96.24%	0.52	14.32	yes
TGT	88.21%	0.51	23.72	yes
TM	96.57%	0.50	25.13	yes
TMPW	93.36%	0.50	22.29	yes
TOM	90.00%	0.54	10.72	yes
TXN	93.93%	0.50	27.36	yes
TXU	95.96%	0.51	19.37	yes
USAI	90.20%	0.50	21.18	yes

Table 10. Continued

Stock	Hit Ratio	Chance	Z-Statistic	Significant
WEN	87.80%	0.51	14.92	yes
WHR	90.97%	0.50	16.74	yes
WIN	83.77%	0.50	10.08	yes
WMT	95.07%	0.50	36.39	yes
WPPGY	97.42%	0.51	17.51	yes
X	92.54%	0.53	18.45	yes
XEL	93.00%	0.50	13.78	yes
XOM	95.08%	0.50	34.19	yes

Table 11. Volume classification training set leave-one-out cross-validation

Stock	Hit Ratio	Chance	Z-Statistic	Significant
AAPL	84.79%	0.50	25.33	yes
ABS	91.71%	0.50	19.37	yes
ADSX	87.66%	0.56	9.88	yes
AEOS	80.29%	0.50	12.40	yes
AET	89.50%	0.51	20.70	yes
AFFX	90.91%	0.51	12.52	yes
ANF	86.23%	0.51	15.34	yes
AOL	94.51%	0.50	45.03	yes
ASKJ	83.84%	0.50	12.25	yes
AXP	89.96%	0.52	29.35	yes
BA	94.70%	0.50	44.58	yes
BAB	94.87%	0.50	28.54	yes
BC	86.32%	0.51	10.39	yes
BDK	80.41%	0.51	10.24	yes
BMY	91.37%	0.50	34.18	yes
BUD	81.78%	0.50	16.76	yes
C	91.81%	0.50	38.73	yes
CCU	90.24%	0.50	19.43	yes
CHIR	92.58%	0.51	17.67	yes
CMTN	78.18%	0.50	8.26	yes
COX	87.34%	0.50	24.54	yes
CREE	91.86%	0.50	12.44	yes
CSCO	92.61%	0.50	39.23	yes
CVC	85.08%	0.51	18.15	yes
DAL	93.28%	0.51	38.62	yes
DD	86.98%	0.50	24.67	yes
DELL	91.03%	0.50	33.33	yes
DOW	86.19%	0.51	20.12	yes
DRI	82.58%	0.53	9.90	yes
EBAY	87.46%	0.50	26.85	yes
ELY	85.01%	0.51	14.91	yes
ERICY	89.32%	0.50	28.47	yes
F	92.21%	0.50	44.12	yes
FDX	87.98%	0.50	24.37	yes
FPL	93.64%	0.51	13.25	yes
GD	91.51%	0.52	26.98	yes
GE	92.63%	0.50	41.03	yes
GLW	89.46%	0.51	26.64	yes
GM	92.89%	0.51	43.67	yes
GPS	84.39%	0.50	23.15	yes
GR	87.45%	0.50	17.30	yes
GTW	89.64%	0.51	24.67	yes
HDI	83.59%	0.51	11.90	yes

Table 11. Continued

Stock	Hit Ratio	Chance	Z-Statistic	Significant
HUM	87.68%	0.50	15.18	yes
IBM	93.09%	0.50	41.05	yes
ILXO	87.32%	0.51	10.54	yes
INTC	91.70%	0.50	40.01	yes
JNJ	91.21%	0.51	29.81	yes
KM	90.80%	0.50	32.39	yes
KR	78.24%	0.52	12.15	yes
LIZ	87.19%	0.50	11.57	yes
LLTC	83.84%	0.50	11.66	yes
LLY	89.91%	0.50	29.59	yes
LMT	92.68%	0.51	32.75	yes
LUV	90.07%	0.50	27.13	yes
MCD	90.75%	0.50	29.74	yes
MO	89.94%	0.51	29.86	yes
MON	90.44%	0.50	19.83	yes
MOT	91.60%	0.50	37.70	yes
MRK	91.98%	0.52	28.52	yes
MYG	85.63%	0.52	19.10	yes
NKE	81.60%	0.50	12.84	yes
NOK	93.13%	0.51	27.28	yes
NSANY	95.29%	0.50	18.67	yes
NVS	95.56%	0.50	24.39	yes
NXTL	88.44%	0.52	20.68	yes
ORCL	95.90%	0.50	31.36	yes
PBG	81.41%	0.50	10.99	yes
PD	95.22%	0.55	13.76	yes
PEP	87.27%	0.50	22.59	yes
PHA	91.53%	0.51	21.84	yes
RBK	79.23%	0.50	10.34	yes
RL	81.41%	0.52	8.38	yes
ROST	78.04%	0.54	7.07	yes
S	87.51%	0.50	23.73	yes
SNE	95.19%	0.50	28.46	yes
SO	91.54%	0.50	13.70	yes
SOI	93.61%	0.52	13.46	yes
TGT	83.42%	0.51	20.68	yes
TM	93.54%	0.50	23.50	yes
TMPW	89.44%	0.50	20.28	yes
TOM	83.18%	0.54	8.69	yes
TXN	90.64%	0.50	25.31	yes
TXU	95.11%	0.51	19.00	yes
USAI	86.31%	0.50	19.13	yes

Table 11. Continued

Stock	Hit Ratio	Chance	Z-Statistic	Significant
WEN	83.49%	0.51	13.16	yes
WHR	88.60%	0.50	15.77	yes
WIN	73.25%	0.50	6.90	yes
WMT	92.09%	0.50	33.97	yes
WPPGY	95.99%	0.51	16.97	yes
X	89.55%	0.53	17.06	yes
XEL	85.60%	0.50	11.41	yes
XOM	95.08%	0.50	34.19	yes

Table 12. Return classification holdout sample results

Stock	Hit Ratio	Press's Q	Significant
AAPL	56.56%	3.81	yes
ABS	48.89%	0.07	no
ADSX	64.04%	8.98	yes
AEOS	63.01%	4.95	yes
AET	50.31%	0.01	no
AFFX	62.07%	1.69	no
ANF	58.33%	1.67	no
AOL	51.58%	0.57	no
ASKJ	37.29%	3.81	no
AXP	56.25%	5.75	yes
BA	49.57%	0.04	no
BAB	42.06%	5.88	no
BC	25.00%	14.00	no
BDK	57.47%	1.94	no
BMY	52.93%	1.52	no
BUD	43.86%	1.72	no
C	42.34%	11.65	no
CCU	51.38%	0.08	no
CHIR	44.23%	0.69	no
CMTN	40.00%	1.00	no
COX	42.11%	2.84	no
CREE	54.08%	0.65	no
CSCO	45.76%	2.80	no
CVC	47.67%	0.19	no
DAL	47.07%	1.46	no
DD	54.31%	1.72	no
DELL	43.91%	4.63	no
DOW	50.00%	0.00	no
DRI	51.95%	0.12	no
EBAY	44.13%	2.93	no
ELY	71.60%	15.12	yes
ERICY	53.45%	1.38	no
F	49.91%	0.00	no
FDX	53.06%	0.92	no
FPL	57.27%	2.33	no
GD	44.62%	2.90	no
GE	47.31%	1.51	no
GLW	55.84%	3.16	yes
GM	54.87%	5.64	yes
GPS	43.67%	2.53	no
GR	46.25%	0.45	no
GTW	45.05%	1.78	no
HDI	35.09%	5.07	no

Table 12. Continued

Stock	Hit Ratio	Press's Q	Significant
HUM	41.67%	1.00	no
IBM	45.38%	2.96	no
ILXO	76.47%	4.76	yes
INTC	48.39%	0.55	no
JNJ	51.45%	0.29	no
KM	48.09%	0.69	no
KR	52.20%	0.31	no
LIZ	37.50%	4.00	no
LLTC	25.58%	10.26	no
LLY	54.51%	1.98	no
LMT	50.13%	0.00	no
LUV	49.28%	0.04	no
MCD	59.35%	9.73	yes
MO	52.13%	0.72	no
MON	51.58%	0.19	no
MOT	45.65%	2.87	no
MRK	51.72%	0.45	no
MYG	51.76%	0.11	no
NKE	60.67%	4.06	yes
NOK	50.00%	0.00	no
NSANY	37.86%	8.26	no
NVS	54.95%	1.78	no
NXTL	48.63%	0.11	no
ORCL	51.33%	0.16	no
PBG	66.67%	7.00	yes
PD	41.79%	1.81	no
PEP	49.68%	0.01	no
PHA	50.24%	0.00	no
RBK	64.41%	4.90	yes
RL	68.42%	2.58	no
ROST	23.53%	4.76	no
S	39.22%	7.12	no
SNE	55.78%	2.66	no
SO	62.22%	5.38	yes
SOI	46.43%	0.14	no
TGT	45.65%	1.04	no
TM	63.19%	11.34	yes
TMPW	48.33%	0.07	no
TOM	68.75%	2.25	no
TXN	57.96%	5.73	yes
TXU	48.33%	0.07	no
USAI	51.85%	0.15	no

Table 12. Continued

Stock	Hit Ratio	Press's Q	Significant
WEN	55.56%	1.11	no
WHR	40.70%	2.98	no
WIN	39.71%	2.88	no
WMT	51.41%	0.25	no
WPPGY	46.38%	0.36	no
X	37.25%	3.31	no
XEL	65.65%	12.83	yes
XOM	51.99%	0.44	no

Table 13. Volume classification holdout sample results

Stock	Hit Ratio	Press's Q	Significant
AAPL	54.38%	1.66	no
ABS	45.26%	1.23	no
ADSX	55.17%	1.24	no
AEOS	66.22%	7.78	yes
AET	54.60%	1.38	no
AFFX	50.00%	0.00	no
ANF	53.73%	0.37	no
AOL	49.11%	0.18	no
ASKJ	45.00%	0.60	no
AXP	45.88%	2.47	no
BA	45.03%	5.47	no
BAB	57.02%	4.63	yes
BC	46.55%	0.28	no
BDK	55.81%	1.16	no
BMY	46.44%	2.21	no
BUD	58.72%	3.31	yes
C	52.02%	0.81	no
CCU	51.75%	0.14	no
CHIR	48.15%	0.07	no
CMTN	20.00%	9.00	no
COX	59.82%	4.32	yes
CREE	44.33%	1.25	no
CSCO	48.68%	0.26	no
CVC	50.59%	0.01	no
DAL	53.30%	1.85	no
DD	56.47%	3.88	yes
DELL	47.13%	1.03	no
DOW	64.49%	8.98	yes
DRI	52.70%	0.22	no
EBAY	56.67%	3.73	yes
ELY	50.62%	0.01	no
ERICY	52.54%	0.71	no
F	51.05%	0.23	no
FDX	56.28%	3.89	yes
FPL	48.21%	0.14	no
GD	56.22%	3.86	yes
GE	51.90%	0.72	no
GLW	47.56%	0.59	no
GM	52.32%	1.25	no
GPS	50.00%	0.00	no
GR	59.26%	2.78	yes
GTW	45.99%	1.20	no
HDI	43.86%	0.86	no

Table 13. Continued

Stock	Hit Ratio	Press's Q	Significant
HUM	47.22%	0.11	no
IBM	50.63%	0.05	no
ILXO	64.71%	1.47	no
INTC	48.15%	0.70	no
JNJ	52.65%	0.95	no
KM	48.72%	0.31	no
KR	44.59%	1.84	no
LIZ	43.75%	1.00	no
LLTC	36.96%	3.13	no
LLY	44.31%	3.19	no
LMT	46.92%	1.48	no
LUV	55.77%	2.77	yes
MCD	49.65%	0.01	no
MO	54.25%	2.89	yes
MON	50.27%	0.01	no
MOT	48.66%	0.27	no
MRK	56.49%	6.23	yes
MYG	40.48%	3.05	no
NKE	42.35%	1.99	no
NOK	47.55%	0.49	no
NSANY	44.00%	1.80	no
NVS	54.29%	1.29	no
NXTL	54.11%	0.99	no
ORCL	50.44%	0.02	no
PBG	43.86%	0.86	no
PD	39.39%	2.97	no
PEP	46.48%	0.70	no
PHA	51.76%	0.25	no
RBK	50.85%	0.02	no
RL	31.25%	2.25	no
ROST	70.59%	2.88	yes
S	51.37%	0.11	no
SNE	61.31%	10.18	yes
SO	51.19%	0.05	no
SOI	34.38%	3.13	no
TGT	46.27%	0.75	no
TM	47.47%	0.41	no
TMPW	44.64%	0.64	no
TOM	37.50%	1.00	no
TXN	53.30%	0.99	no
TXU	36.51%	4.59	no
USAI	45.95%	0.73	no

Table 13. Continued

Stock	Hit Ratio	Press's Q	Significant
WEN	56.18%	1.36	no
WHR	49.44%	0.01	no
WIN	38.57%	3.66	no
WMT	48.59%	0.25	no
WPPGY	56.34%	1.14	no
X	34.38%	6.25	no
XEL	54.89%	1.27	no
XOM	59.36%	9.93	yes

Table 14. Return classification with x_1 training set results

Stock	Step Entered	Hit Ratio	Chance	Z-Statistic	Significant
AAPL	30	93.88%	0.50	32.13	yes
ABS	94	95.27%	0.50	20.81	yes
ADSX	25	89.96%	0.51	11.82	yes
AEOS	2	87.98%	0.50	15.37	yes
AET	38	93.98%	0.51	22.73	yes
AFFX	45	97.08%	0.52	14.12	yes
ANF	27	88.09%	0.54	14.80	yes
AOL	11	97.16%	0.51	47.03	yes
ASKJ	26	91.38%	0.51	14.61	yes
AXP	6	95.92%	0.51	34.98	yes
BA	619	96.81%	0.50	46.78	yes
BAB	77	96.63%	0.50	29.59	yes
BC	25	91.26%	0.51	11.50	yes
BDK	1	84.69%	0.50	11.78	yes
BMY	1	95.24%	0.50	37.74	yes
BUD	1	85.74%	0.50	18.71	yes
C	10	96.01%	0.50	42.79	yes
CHIR	65	93.68%	0.50	18.27	yes
CMTN	13	87.44%	0.50	10.98	yes
COX	20	91.13%	0.50	27.00	yes
CVC	44	89.67%	0.50	20.79	yes
DAL	2	96.95%	0.50	43.00	yes
DD	39	95.61%	0.51	29.37	yes
DELL	108	94.19%	0.50	36.22	yes
DOW	3	94.10%	0.50	24.81	yes
DRI	52	92.53%	0.57	11.95	yes
EBAY	1	96.52%	0.50	33.05	yes
ELY	77	93.11%	0.50	18.26	yes
ERICY	1	94.53%	0.51	31.43	yes
F	5	95.55%	0.50	47.46	yes
FDX	36	92.08%	0.51	25.90	yes
GD	4	94.05%	0.50	29.56	yes
GE	295	93.40%	0.50	42.04	yes
GLW	16	94.48%	0.51	30.01	yes
GM	113	95.62%	0.50	47.34	yes
GPS	18	92.12%	0.52	26.88	yes
GR	7	92.59%	0.50	19.49	yes
GTW	95	92.99%	0.51	27.23	yes
HDI	4	89.14%	0.50	13.75	yes
HUM	15	94.40%	0.53	16.43	yes
IBM	3	95.47%	0.51	43.49	yes
INTC	5	95.47%	0.51	43.10	yes
JNJ	1	95.89%	0.50	33.55	yes

Table 14. Continued

Stock	Step Entered	Hit Ratio	Chance	Z-Statistic	Significant
KM	1	92.68%	0.50	33.93	yes
KR	14	86.54%	0.50	16.76	yes
LIZ	40	90.24%	0.52	12.13	yes
LLTC	1	92.23%	0.51	14.31	yes
LLY	37	95.61%	0.50	33.31	yes
LMT	168	96.23%	0.50	35.92	yes
LUV	1	94.85%	0.51	29.69	yes
MCD	4	93.27%	0.50	32.11	yes
MO	216	94.13%	0.51	33.29	yes
MON	10	93.58%	0.51	20.89	yes
MOT	12	94.56%	0.51	39.62	yes
MRK	1	96.03%	0.50	32.57	yes
MYG	52	84.43%	0.50	19.57	yes
NKE	1	91.57%	0.50	16.93	yes
NOK	3	94.87%	0.50	29.07	yes
NSANY	1	96.76%	0.51	19.18	yes
NVS	43	97.78%	0.50	25.67	yes
NXTL	18	91.46%	0.50	23.89	yes
PBG	33	82.86%	0.51	11.33	yes
PD	24	97.28%	0.56	14.25	yes
PEP	12	90.67%	0.52	23.85	yes
PHA	16	95.66%	0.50	24.67	yes
RBK	9	86.60%	0.51	12.93	yes
ROST	10	86.11%	0.52	9.90	yes
S	55	91.77%	0.55	23.55	yes
SNE	2	97.22%	0.51	29.50	yes
TGT	30	92.28%	0.53	25.18	yes
TM	88	97.54%	0.50	25.72	yes
TOM	2	92.00%	0.51	12.44	yes
TXN	205	95.80%	0.50	28.54	yes
TXU	19	96.88%	0.51	20.00	yes
USAI	1	93.92%	0.50	23.18	yes
WEN	34	90.65%	0.53	15.76	yes
WHR	11	96.46%	0.50	19.00	yes
WIN	23	89.91%	0.51	11.76	yes
WMT	1	97.07%	0.50	38.19	yes
X	72	89.62%	0.55	16.52	yes
XEL	48	90.97%	0.50	13.61	yes
XOM	3	95.45%	0.50	34.87	yes

Table 15. Volume classification with x_1 training set results

Stock	Step X1 Entered	Hit Ratio	Chance	Z-Statistic	Significant
AAPL	1	93.60%	0.50	31.55	yes
ABS	1	98.28%	0.50	22.05	yes
ADSX	2	90.27%	0.55	10.71	yes
AEOS	1	87.47%	0.50	15.23	yes
AET	1	96.73%	0.51	24.37	yes
AFFX	5	98.31%	0.51	14.58	yes
ANF	1	91.67%	0.51	17.63	yes
AOL	1	97.57%	0.50	47.65	yes
ASKJ	1	95.99%	0.50	16.54	yes
AXP	1	96.51%	0.52	34.37	yes
BA	10	97.70%	0.50	47.11	yes
BAB	1	97.69%	0.50	30.05	yes
BC	2	94.61%	0.50	12.65	yes
BDK	1	91.16%	0.51	13.91	yes
BMY	1	96.75%	0.50	38.44	yes
BUD	1	88.00%	0.50	19.72	yes
C	1	96.75%	0.50	42.87	yes
CCU	1	92.88%	0.50	20.22	yes
CHIR	1	96.15%	0.51	19.12	yes
CMTN	1	90.23%	0.50	11.74	yes
COX	2	94.47%	0.50	29.00	yes
CSCO	1	96.43%	0.50	42.25	yes
CVC	1	89.41%	0.51	20.25	yes
DAL	1	97.66%	0.51	42.36	yes
DD	1	93.82%	0.50	28.82	yes
DELL	1	97.34%	0.50	38.08	yes
DOW	1	91.28%	0.52	22.22	yes
DRI	6	91.81%	0.54	12.67	yes
EBAY	1	95.45%	0.50	32.16	yes
ELY	1	90.34%	0.51	16.62	yes
ERICY	1	94.84%	0.50	32.07	yes
F	1	96.72%	0.50	48.38	yes
FDX	1	94.17%	0.50	27.85	yes
FPL	2	95.32%	0.51	13.72	yes
GD	1	96.16%	0.52	29.89	yes
GE	1	96.09%	0.50	44.18	yes
GLW	1	94.59%	0.51	30.11	yes
GM	1	97.34%	0.51	47.84	yes
GPS	1	92.51%	0.50	28.46	yes
GR	20	91.94%	0.50	19.14	yes
GTW	1	96.09%	0.51	28.68	yes
HDI	1	92.88%	0.50	14.99	yes
HUM	3	91.84%	0.50	16.56	yes

Table 15. Continued

Stock	Step X1 Entered	Hit Ratio	Chance	Z-Statistic	Significant
IBM	1	96.97%	0.51	44.31	yes
ILXO	6	93.07%	0.50	12.10	yes
INTC	1	95.82%	0.50	43.22	yes
JNJ	1	96.14%	0.51	33.16	yes
KM	15	94.46%	0.50	34.96	yes
KR	1	87.76%	0.52	16.36	yes
LIZ	2	90.68%	0.50	12.50	yes
LLTC	9	90.14%	0.50	13.76	yes
LLY	1	93.06%	0.50	31.35	yes
LMT	1	96.65%	0.51	35.41	yes
LUV	1	94.57%	0.50	29.88	yes
MCD	8	94.39%	0.50	32.14	yes
MO	1	96.04%	0.51	33.94	yes
MON	1	96.43%	0.50	22.49	yes
MOT	1	96.12%	0.50	41.26	yes
MRK	1	96.77%	0.52	31.65	yes
MYG	1	91.57%	0.52	22.48	yes
NKE	1	93.70%	0.50	17.76	yes
NOK	3	95.65%	0.51	28.90	yes
NSANY	1	97.88%	0.50	19.73	yes
NVS	1	98.06%	0.50	25.73	yes
NXTL	92	91.76%	0.52	22.57	yes
ORCL	1	98.29%	0.50	33.00	yes
PBG	1	83.97%	0.50	11.90	yes
PD	22	97.27%	0.55	14.46	yes
PEP	1	95.65%	0.50	27.67	yes
PHA	1	95.49%	0.51	23.99	yes
RBK	3	88.50%	0.50	13.62	yes
ROST	2	85.98%	0.54	9.41	yes
S	1	94.91%	0.50	28.40	yes
SNE	1	98.00%	0.50	30.23	yes
SO	1	95.22%	0.50	14.92	yes
SOI	56	96.24%	0.52	14.32	yes
TGT	1	91.11%	0.51	25.55	yes
TM	1	96.57%	0.50	25.13	yes
TMPW	1	96.83%	0.50	24.08	yes
TOM	9	92.73%	0.54	11.53	yes
TXN	1	94.44%	0.50	27.68	yes
TXU	4	94.68%	0.51	18.81	yes
USAI	1	93.80%	0.50	23.08	yes
WEN	1	93.30%	0.51	17.17	yes
WHR	21	91.92%	0.50	17.13	yes

Table 15. Continued

Stock	Step X1 Entered	Hit Ratio	Chance	Z-Statistic	Significant
WIN	1	88.16%	0.50	11.41	yes
WMT	1	97.93%	0.50	38.71	yes
WPPGY	1	98.57%	0.51	17.94	yes
X	1	94.03%	0.53	19.14	yes
XEL	4	94.94%	0.50	14.41	yes
XOM	1	95.98%	0.50	34.87	yes

Table 16. Return classification with x_1 holdout sample results

Stock	Hit Ratio	Press's Q	Significant
AAPL	57.01%	4.35	yes
ABS	54.81%	1.25	no
ADSX	56.14%	1.72	no
AEOS	49.32%	0.01	no
AET	49.07%	0.06	no
AFFX	41.38%	0.86	no
ANF	61.67%	3.27	yes
AOL	52.98%	2.03	no
ASKJ	44.07%	0.83	no
AXP	51.90%	0.53	no
BA	50.43%	0.04	no
BAB	47.21%	0.73	no
BC	35.71%	4.57	no
BDK	50.57%	0.01	no
BMY	46.40%	2.31	no
BUD	45.61%	0.88	no
C	46.77%	2.06	no
CHIR	61.54%	2.77	yes
CMTN	40.00%	1.00	no
COX	35.09%	10.14	no
CVC	55.81%	1.16	no
DAL	49.65%	0.02	no
DD	55.17%	2.48	no
DELL	41.03%	10.05	no
DOW	50.00%	0.00	no
DRI	57.14%	1.57	no
EBAY	46.95%	0.79	no
ELY	60.49%	3.57	yes
ERICY	50.69%	0.06	no
F	56.72%	9.81	yes
FDX	49.80%	0.00	no
GD	47.81%	0.48	no
GE	42.31%	12.31	no
GLW	51.52%	0.21	no
GM	55.87%	8.22	yes
GPS	38.61%	8.20	no
GR	37.50%	5.00	no
GTW	48.35%	0.20	no
HDI	50.88%	0.02	no
HUM	41.67%	1.00	no
IBM	47.40%	0.94	no
INTC	51.04%	0.23	no
JNJ	52.03%	0.57	no

Table 16. Continued

Stock	Hit Ratio	Press's Q	Significant
KM	46.82%	1.91	no
KR	50.94%	0.06	no
LIZ	43.75%	1.00	no
LLTC	37.21%	2.81	no
LLY	48.77%	0.15	no
LMT	49.62%	0.02	no
LUV	43.54%	3.49	no
MCD	54.68%	2.43	no
MO	49.62%	0.02	no
MON	53.68%	1.03	no
MOT	45.38%	3.23	no
MRK	51.19%	0.21	no
MYG	48.24%	0.11	no
NKE	30.34%	13.76	no
NOK	50.00%	0.00	no
NSANY	47.14%	0.46	no
NVS	64.29%	14.86	yes
NXTL	54.79%	1.34	no
PBG	68.25%	8.40	yes
PD	35.82%	5.39	no
PEP	47.77%	0.31	no
PHA	40.00%	8.20	no
RBK	59.32%	2.05	no
ROST	58.82%	0.53	no
S	51.63%	0.16	no
SNE	61.81%	11.10	yes
TGT	55.07%	1.42	no
TM	58.28%	4.47	yes
TOM	75.00%	4.00	yes
TXN	57.52%	5.12	yes
TXU	50.00%	0.00	no
USAI	51.85%	0.15	no
WEN	60.00%	3.60	yes
WHR	34.88%	7.86	no
WIN	27.94%	13.24	no
WMT	45.14%	3.01	no
X	47.06%	0.18	no
XEL	64.89%	11.61	yes
XOM	55.23%	3.04	yes

Table 17. Volume classification with x_1 holdout sample results

Stock	Hit Ratio	Press's Q	Significant
AAPL	58.99%	7.01	yes
ABS	58.21%	3.61	yes
ADSX	62.28%	6.88	yes
AEOS	46.58%	0.34	no
AET	49.33%	0.03	no
AFFX	68.97%	4.17	yes
ANF	56.67%	1.07	no
AOL	53.15%	2.21	no
ASKJ	47.46%	0.15	no
AXP	46.24%	2.03	no
BA	50.74%	0.12	no
BAB	57.21%	4.76	yes
BC	73.21%	12.07	yes
BDK	57.65%	1.99	no
BMY	50.23%	0.01	no
BUD	62.26%	6.38	yes
C	48.02%	0.75	no
CCU	54.63%	0.93	no
CHIR	52.94%	0.18	no
CMTN	52.00%	0.04	no
COX	64.55%	9.31	yes
CSCO	43.73%	5.89	no
CVC	59.04%	2.71	yes
DAL	52.35%	0.89	no
DD	52.38%	0.52	no
DELL	58.31%	8.47	yes
DOW	63.54%	7.04	yes
DRI	48.65%	0.05	no
EBAY	60.58%	9.31	yes
ELY	55.56%	1.00	no
ERICY	55.15%	2.88	yes
F	56.45%	8.51	yes
FDX	54.73%	2.18	no
FPL	41.82%	2.95	no
GD	53.82%	1.45	no
GE	45.25%	4.46	no
GLW	45.45%	1.91	no
GM	44.35%	7.35	no
GPS	56.96%	3.06	yes
GR	57.50%	1.80	no
GTW	60.44%	7.93	yes
HDI	52.63%	0.16	no
HUM	36.11%	2.78	no

Table 17. Continued

Stock	Hit Ratio	Press's Q	Significant
IBM	50.31%	0.01	no
ILXO	41.18%	0.53	no
INTC	53.75%	2.85	yes
JNJ	54.19%	2.35	no
KM	48.71%	0.31	no
KR	57.96%	3.98	yes
LIZ	45.16%	0.58	no
LLTC	44.19%	0.58	no
LLY	38.59%	12.55	no
LMT	54.29%	2.83	yes
LUV	57.84%	5.02	yes
MCD	47.46%	0.71	no
MO	51.54%	0.37	no
MON	56.04%	2.66	no
MOT	59.57%	13.59	yes
MRK	62.40%	22.56	yes
MYG	43.37%	1.46	no
NKE	59.52%	3.05	yes
NOK	46.77%	0.84	no
NSANY	56.67%	2.13	no
NVS	59.17%	5.69	yes
NXTL	62.07%	8.45	yes
ORCL	53.24%	0.91	no
PBG	63.16%	3.95	yes
PD	53.85%	0.38	no
PEP	56.34%	2.28	no
PHA	58.76%	5.96	yes
RBK	51.72%	0.07	no
ROST	76.47%	4.76	yes
S	52.08%	0.25	no
SNE	59.38%	6.75	yes
SO	48.78%	0.05	no
SOI	40.00%	1.00	no
TGT	39.85%	5.48	no
TM	50.97%	0.06	no
TMPW	42.86%	1.14	no
TOM	31.25%	2.25	no
TXN	60.71%	10.29	yes
TXU	46.67%	0.27	no
USAI	49.04%	0.04	no
WEN	58.62%	2.59	no
WHR	50.00%	0.00	no

Table 17. Continued

Stock	Hit Ratio	Press's Q	Significant
WIN	45.45%	0.55	no
WMT	63.75%	23.38	yes
WPPGY	39.13%	3.26	no
X	33.33%	5.67	no
XEL	62.31%	7.88	yes
XOM	51.14%	0.14	no

CHAPTER 6
DISCUSSION, CONCLUSIONS, AND FUTURE DIRECTIONS

The purpose of Chapter 6 is to summarize the findings in this dissertation in light of the proposed research questions, conclude and provide directions for future research. Section 6.1 contains the summary. Section 6.2 provides a discussion of the results. Section 6.3 provides a conclusion. Finally, Section 6.4 discusses the directions for future research.

6.1 Summary

The purpose of this dissertation is to provide a methodology for an organization to use to aid in environmental scanning of web documents. The methodology proposed involves combining the vector space model representation of the documents with linear discriminant analysis as the method of classification of the documents. The vector space model is used to represent documents in the text classification literature and linear discriminant analysis is a well-founded method of classification. Linear discriminant analysis provides a linear discriminant function found via a training set of documents with known classification. The linear discriminant function can then be used to classify future documents that appear on the web to give an organization an idea about what new documents are indicating. The process of collecting, representing and classifying the training set and the new documents is automated via Java programs.

The methodology developed in this dissertation is tested empirically on news documents that appear at designated web sites about a set of 186 publicly traded

companies. After excluding companies that have too few documents to analyze, incomplete document collections and discontinued public trading, a set of 93 stocks is used in this empirical study. The study is conducted in light of four research questions outlined in Section 3.2 and Section 4.2. The results are organized according to these four questions.

The first research question addresses how well the linear discriminant function classifies the training set of documents. In order to determine the answer to this question the validity of the 80% training set is checked using the z-statistic comparing the hit ratio to the proportional chance criteria for both the training sample classification matrix and the leave-one-out classification matrix. The proportional chance criteria value (chance) is given for each stock with classification based on return in Table 8. Additionally, Table 8 provides the hit ratio and z-statistic based on the training sample classification matrix. Table 9 provides the same information for the leave-one-out classification matrix. For both tables, for every stock the z-statistic comparing chance classification to the classification matrices is significant at 10%. Table 10 provides the proportional chance criteria value (chance), the hit ratio and z-statistic for each stock with classification based on volume. Table 11 provides the same information for the leave-one-out classification matrix. Once again, for both tables, for every stock the z-statistic is significant at 10%. The average hit ratio for return classification for the training classification matrix is 92.11% and for the leave-one-out classification matrix is 89.06%. The average hit ratio for volume classification for the training classification matrix is 91.68% and for the leave-one-out classification matrix is 88.68%. Based on the evidence provided, the linear discriminant function does very well in classifying the training set of documents.

The second research question addresses how well the linear discriminant function derived from the training set of documents does on classifying the 20% holdout sample. Specifically, does the linear discriminant function predict the correct classification in the holdout sample better than random guessing? To address this question, each document in the holdout sample is classified according to the linear discriminant function and Press's Q is computed based on the number of correctly classified documents, the total number of documents and the number of classification groups using a chi-square distribution with one degree of freedom for two-group classification. Using return as the classification mechanism, 16 out of 93 or 17.20% of the stocks had holdout classification matrices that were statistically significant at 10%. Using volume as the classification mechanism, the result is exactly the same. Therefore, the methodology does a fairly good job of classification of the holdout sample for both classification mechanisms.

Research question three is used to determine if there is a noticeable difference in the performance of the two classification mechanisms. Based on the classification results for the training set and the holdout sample, there is not a noticeable difference in the performance of the two classification mechanisms.

Finally, the fourth research question addresses whether adding an independent variable, x_1, to the set of variables that can potentially enter the linear discriminant model via the stepwise discriminant procedure improves the prediction accuracy in the holdout sample. The variable x_1 is calculated based on the stock's classification in the three days prior to the current classification date. For both classification mechanisms, the variable x_1 enters the linear discriminant model for most of the stocks. For return classification, the variable enters the model for 82 out of 93 stocks. For volume classification, the

variable enters the model for 91 out of 93 stocks. However, for volume classification the step at which the variable enters the model on average, 4.12 is much earlier than for return classification, 41.48. Out of the 82 stocks with x_1 entering the model in the stepwise discriminant step, 15 have holdout samples with classification matrices that are significant at 10% using Press's Q. For the same set of 82 stocks, 14 have holdout samples with classification matrices that are significant without adding the variable x_1 to the set of independent variables. There is not a significant difference in the prediction accuracy when adding x_1 to the model with return classification. Out of the 91 stocks with x_1 entering the model in the stepwise discriminant step, 32 have holdout samples with classification matrices that are significant at 10% using Press's Q. For the same set of 91 stocks, 16 have holdout samples with classification matrices that are significant without adding the variable x_1 to the set of independent variables. The p-value for the difference in the two proportions, 32 out of 91 versus 16 out of 91, is 0.0036. There is a very significant difference in the prediction accuracy when adding x_1 to the model with volume classification.

6.2 Discussion

In this study the relationship between news and stock returns and between news and trading volume is examined. The discriminant function computed using the terms appearing in news articles in the training set as the independent variables and classification based on stock returns as the dependent variable in the discriminant analysis procedure has predictive capability for the holdout sample in 16 out of 93 stocks. Based on this predictive capability for these 16 stocks, a profitable daily trading strategy for these stocks could be implemented. The ability to develop such a strategy contradicts

the efficient market hypothesis (Fama 1970). Additionally, the predictive capability of the results shows a link between news and stock returns. The addition of an independent variable that represents the stock's classification based on return for the three days prior to classification does not provide a significant increase in the number of stocks in the holdout sample with classification matrices that are significant. However, the set of significant stocks changed considerably, with 10 different stocks appearing in the set of stocks with significant holdout classification matrices.

The discriminant function computed using the terms appearing in news articles in the training set as the independent variables and classification based on daily changed in trading volume as the dependent variable in the discriminant analysis procedure has predictive capability for the holdout sample in 16 out of 93 stocks. The predictive capability for the 16 stocks as well as the training set classification accuracy for all 93 stocks illustrates that there is a link between news and subsequent trading volume. This link between news and subsequent trading volume is consistent with financial literature. In this study, classification is based on daily changes in trading volume, as opposed to levels of trading activity or turnover ratio, the measurement typically used in financial studies. Based on the results of this study, an interesting application would be to determine which terms signal a decrease in trading activity and which signal an increase. The addition of an independent variable that represents the stock's classification based on trading volume for the three days prior to classification does provide a significant increase in the number of stocks, 32, in the holdout sample with classification matrices that are significant. The number of stocks that include the new independent variable in

the linear discriminant model via the stepwise linear discriminant procedure is very high at 91 out of 93 stocks.

In summary, the addition of an independent variable that represents prior performance of the stock results in a higher percent of stocks with the new independent variable entering in the stepwise procedure for volume classification than return classification. The new variable also results in a significantly lower average entering step for the new variable in the stepwise procedure for volume classification than return classification. Finally, the new variable results in a higher number of stocks with significant holdout classification matrices for volume classification than return classification. These differences in volume and return classification indicate that the previous volume activity is a significant aspect of the linear discriminant model as compared to previous stock returns.

6.3 Conclusion

In conclusion, the environmental scanning methodology developed in this dissertation is automated, well founded and useful to an organization. The methodology is validated empirically. The training set has excellent classification results, with 100% of the stocks' training classification matrices and leave-one-out classification matrices having statistical significance using either volume or return as the classification mechanism. The predictive capability of the linear discriminant function, calculated using the training set, shows great promise with 17% of the stocks having holdout classification matrices with statistical significance. Adding the independent variable x_1 to the set of potential variables to enter the linear discriminant model via the stepwise discriminant procedure improves the percentage of stocks with statistical significance in

the holdout classification matrices to 35% when classification is based on trading volume. Finally, the methodology has great versatility as shown by the ease of incorporating a variety of independent variables, the capability of handling large document collections with a large number of terms and the adaptability to a variety of applications.

6.4 Future Directions

Based on the environmental scanning methodology, there are several directions that this current research could go. The volume classification mechanism can be changed to classify according to level of trading volume or trading turnover, defined as the number of shares traded divided by the number of shares outstanding, as opposed to change in trading volume as compared to the previous day. The impact of a variable that represents the nature and reliability of the source of a new article can be investigated. A trading strategy can be developed and tested based on the stocks that show prediction capability in the holdout sample with return classification. Linear programming approaches to the discriminant analysis problem should be investigated, as they are not hindered by violations in the assumptions made in the Fisher (1936) approach.

Additionally, the environmental scanning methodology developed in this dissertation can be used to analyze other areas. One additional application area involves the monitoring of the content of stock chat-rooms or message boards with the same classification mechanisms discussed in this dissertation. Non-financial applications are easily imagined too. For example, chat room discussions, critic reviews and press releases for movies are documents that can be classified according to the success of a

movie. As new movies are released, their success can be predicted based on the content of the aforementioned documents.

APPENDIX A
EXAMPLE

Let D be a collection of three documents d_1, d_2, and d_3. The documents in the space D are classified according to three terms t_1, t_2, and t_3. Let t_1, t_2, t_3, d_1, d_2, d_3 and D be defined as follows.

$$t_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad t_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \textit{and} \quad t_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$D = \begin{bmatrix} 5 & 0 & 2 \\ 1 & 3 & 3 \\ 4 & 0 & 7 \end{bmatrix}$$

$$d_1 = 5t_1 + 2t_3$$

$$d_2 = t_1 + 3t_2 + 3t_3$$

$$d_3 = 4t_1 + 7t_3$$

Consider a query $q_1 = (0 \quad 1 \quad 2)$ *or equivalently* $q_1 = t_2 + 2t_3$ then

$$\begin{aligned} d_1q_1 &= (5t_1 + 2t_3)(t_2 + 2t_3) \\ &= 5t_1t_2 + 2t_2t_3 + 10t_1t_3 + 4t_3t_3 \\ &= (5*0) + (2*0) + (10*0) + (4*1) \\ &= 4 \end{aligned}$$

$$\begin{aligned} d_2q_1 &= (t_1 + 3t_2 + 3t_3)(t_2 + 2t_3) \\ &= t_1t_2 + 2t_1t_3 + 3t_2t_2 + 9t_2t_3 + 6t_3t_3 \\ &= (1*0) + (2*0) + (3*1) + (9*0) + (6*1) \\ &= 9 \end{aligned}$$

$$\begin{aligned} d_3q_1 &= (4t_1 + 7t_3)(t_2 + 2t_3) \\ &= 4t_1t_2 + 8t_1t_3 + 7t_2t_3 + 14t_3t_3 \\ &= (4*0) + (8*0) + (7*0) + (14*1) \\ &= 14 \end{aligned}$$

Hence, d_3 has the highest similarity to q_1.

APPENDIX B
WEB SITES, STOCK SYMBOLS AND STOPWORDS

Table 18. Financial web sites

URL of Web Site	Description
http://smartmoney.com	SmartMoney financial web site
http://finance.yahoo.com	Yahoo financial web site
http://morningstar.com	Morningstar financial web site
http://tradetrek.com	Tradetrek financial web site

Table 19. Stocks

Stock Symbol	Name
AAPL	Apple Computer Inc
ABS	Albertson's Inc
ACNAF	Air Canada
ADSX	Applied Digital Solutions Inc
AEOS	American Eagle Outfitters Inc
AET	Aetna Inc
AFFX	Affymetrix Inc
AIMM	AutoImmune Inc
ALO	Alpharma Inc
ALR	Allied Research Corp
ANF	Abercrombie & Fitch Co Retail (Apparel)
AOL	AOL Time Warner Inc
ARNA	Arena Pharmaceutical Inc
ASCA	Ameristar Casinos Inc
ASF	Administaff Inc
ASKJ	Ask Jeeves Inc
ATAC	Aftermarket Technology Corp
AVIR	Aviron
AXP	American Express Co
BA	The Boeing Co
BAB	British Airways PLC
BAS	
BC	Brunswick Corp Recreational Products
BDK	Black & Decker Corp
BMY	Bristol-Myers Squibb Co Major Drugs

Table 19. Continued

Stock Symbol	Name
BNG	Benetton Group SpA
BOSA	Boston Acoustics Inc
BTU	Peabody Energy Corp
BUD	Anheuser-Busch Companies Inc
BUR	Benetton Group SpA
BVF	Biovail Corp
C	Citigroup
CBIZ	Century Business Services Inc
CCMP	Cabot Microelectronics Corp
CCU	Clear Channel Communications Inc
CHDN	Churchill Downs Inc
CHIR	Chiron Corp
CMTN	Copper Mountain Networks Inc
COKE	Coca-Cola Bottling Co Consolidated
COX	Cox Communications Inc
CREE	Cree Inc
CRV	Coast Distribution System (DE)
CSCO	Cisco
CTLM	Centillium Communications Inc
CVC	Cablevision Systems Corp
CYBR	Cyber-Care Inc.
DAL	Delta Air Lines Inc
DD	E.I. Du Pont De Nemours & Co (DuPont)
DELL	Dell Computer Corp
DNDN	Dendreon Corp
DOW	Dow Chemical Co
DPMI	Dupont Photomasks Inc
DRI	Darden Restaurants Inc
EBAY	eBay Inc
ELY	Callaway Golf Co
ENDP	Endo Pharmaceuticals Holdings Inc
ERICY	Major Drugs Ericsson, Telefonab. L M AB
ERJ	Embraer SA
EW	Edwards Lifesciences Corp
F	Ford Motor Co
FA	Fairchild Corp
FDX	Fedex Corp (Federal Express)
FOSL	Fossil Inc
FPL	Florida Power & Light Co
FRK	Florida Rock Industries Inc

Table 19. Continued

Stock Symbol	Name
FRX	Forest Laboratories Inc
GD	General Dynamics Corp
GE	General Electric Comp.
GES	Guess ? Inc
GLCBY	Globo Cabo SA
GLK	Great Lakes Chemical Corp
GLW	Corning Inc
GM	General Motors Corp
GOSHA	Oshkosh B'Gosh Inc
GPS	Gap Inc
GR	Goodrich Corp
GTW	Gateway Inc
HDI	Harley Davidson
HELE	Helen of Troy Corp
HGGR	Haggar Corp
HLYW	Hollywood Entertainment
HOLL	Hollywood Media Corp
HTVN	Hispanic Television Network Inc
HUM	Humana Inc
HYDL	Hydril Co.
IBM	International Business Machines Corp
ILXO	ILEX Oncology Inc
INTC	Intel Corp
JNJ	Johnson & Johnson Inc Major Drugs
JUNI	Juniper Group Inc
KELYA	Kelly Services Inc
KM	K Mart Corp
KNBWY	Kirin Brewery Co Ltd
KR	Kroger Co
LIZ	Liz Claiborne Inc
LLTC	Linear Technology Corp
LLY	Eli Lilly and Co Major Drugs
LMGA	AT&T Corp
LMT	Lockheed Martin Corp
LTRE	Learning Tree
LUV	Southwest Airlines Inc
LZB	La-Z-Boy Inc
MCD	McDonald's Corp
MCO	Moody's Corp
MO	Philip Morris Companies Inc
MON	Monsanto Co
MOND	Robert Mondavi Corp

Table 19. Continued

Stock Symbol	Name
MOT	Motorola Inc
MOV	Movado Group Inc Jewelry & Silverware
MPPP	MP3.com Inc Recreational Products
MRK	Merck & Co Inc Major Drugs
MSFT	Microsoft
MSN	Emerson Radio Corp
MTIC	MTI Technology Corp
MTIC	MTI Technology Corp.
MUEI	Micron Electronics Inc
MVL	Marvel Enterprises, Inc
MYG	Maytag Corp
NKE	Nike Inc
NOK	Nokia
NSANY	Nissan Motor Co Ltd
NSAT	Norsat Intl Inc.
NVS	Novartis AG Major Drugs
NXTL	Nextel Communications
OCQ	Oneida Ltd Jewelry & Silverware
ORAL	OrthAlliance Inc
ORCL	Oracle Corp.
OSI	Outback Steakhouse Inc
PB	Panamerican Beverages Inc
PBG	The Pepsi Bottling Group Inc
PD	Phelps Dodge Corp Metal Mining
PEGS	Pegasus Solutions Inc
PEP	Pepsico Inc
PERY	Perry Ellis International Inc
PHA	Pharmacia Corp
PIK	Water Pik Technologies Inc
PNW	Pinnacle West Capital Corp Electric Utilities
PNY	Piedmont Natural Gas Inc
POM	Potomac Electric Power Co
RBK	Reebok International Ltd Footwear
RGB	RG Barry Corp Footwear
RL	Polo Ralph Lauren Corp
ROFO	Rockford Corp
ROST	Ross Stores Inc
RS	Reliance Steel And Aluminum Co
RST	Boca Resorts Inc.
RVWD	Ravenswood Winery Inc
S	Sears, Roebuck and Co
SABI	Swiss Army Brands Inc

Table 19. Continued

Stock Symbol	Name
SAMC	Samsonite Corp
SBTV	SBS Broadcasting SA
SHFL	Shuffle Master Inc
SHW	Sherwin-Williams Co
SLAB	Hollyewood Entertainment
SLVN	Sylvan Learning Systems Inc
SNE	Sony Corp
SO	Southern Company Inc Electric Utilities
SOI	Solutia Inc
SPAR	Spartan Motors Inc
SWC	Stillwater Mining Co
TBL	Timberland Co
TGT	Target Corp
TM	Toyota Motor Comp.
TMPW	TMP Worldwide Inc
TOM	Tommy Hilfiger Corp
TREE	LendingTree Inc
TUP	Tupperware Corporation
TXN	Texas Instruments Incorporated
TXU	TXU Corp
UBET	Youbet.com Inc
USAI	USA Networks Inc
USON	US Oncology Inc
VNGD	Vanguard Airlines Inc
WEC	Wisconsin Energy Corp Electric Utilities
WEN	Wendy's International Inc
WHR	Whirlpool Corp
WIN	Winn-Dixie Stores Inc
WLDA	World Airways Inc
WMT	Wal-Mart Stores Inc
WOR	Worthington Industries Inc
WPPGY	WPP Group PLC
WS	Weirton Steel Corp Iron & Steel
X	USX-US Steel Group Iron & Steel
XEL	Xcel Energy Inc
XOM	Exxon Mobil Corp.
YSTM	YouthStream Media Networks Inc
ZOOX	Gadzoox Networks Inc

Table 20. Stopword list

Stopword List				
a	ends	knew	per	thoughts
about	enough	know	perhaps	three
above	etc	known	place	through
according	even	knows	places	throughout
across	evenly	l	point	thru
actually	ever	large	pointed	thus
adj	every	largely	pointing	to
after	everybody	last	points	today
afterwards	everyone	later	possible	together
again	everything	latest	present	too
against	everywhere	latter	presented	took
all	except	latterly	presenting	toward
almost	f	least	presents	towards
alone	face	less	problem	trillion
along	faces	let	problems	turn
already	fact	lets	put	turned
also	facts	let's	puts	turning
although	far	like	r	turns
always	felt	likely	rather	twenty
among	few	long	really	two
amongst	find	longer	recent	u
an	finds	longest	recently	under
and	fifty	ltd	right	unless
another	first	m	room	unlike
any	five	made	rooms	unlikely
anybody	for	make	s	until
anyhow	former	making	said	up
anyone	formerly	makes	same	upon
anything	forty	man	saw	us
anywhere	found	many	say	use
are	four	may	says	used
areas	from	maybe	second	uses
area	full	me	seconds	using
aren't	fully	meantime	see	v
around	further	meanwhile	seem	very
as	furthered	member	seemed	via
ask	furthering	members	seeming	w
asked	furthers	men	seems	want
asking	g	might	sees	wanted
asks	gave	million	seven	wanting
at	general	miss	seventy	wants
away	generally	more	several	was
b	get	moreover	shall	wasn't

Table 20. Continued

Stopword List				
back	gets	most	she	way
backed	give	mostly	she'd	ways
backing	given	mr	she'll	we
backs	gives	mrs	she's	we'd
be	go	much	should	we'll
became	going	must	shouldn't	we're
because	good	my	show	we've
become	goods	myself	showed	well
becomes	got	n	showing	wells
becoming	great	namely	shows	went
been	greater	necessary	side	were
before	greatest	need	sides	weren't
beforehand	group	needed	since	what
began	grouped	needing	six	what'll
begin	grouping	needs	sixty	what's
beginning	groups	neither	small	what've
behind	h	never	smaller	whatever
being	had	nevertheless	smallest	when
beings	has	new	so	whence
below	hasn't	newer	some	whenever
beside	have	newest	somebody	where
besides	having	next	somehow	where's
best	haven't	non	someone	whereafter
better	having	nine	something	whereas
between	he	ninety	sometime	whereby
beyond	he'd	no	sometimes	wherein
big	he'll	nobody	somewhere	whereupon
billion	he's	none	state	wherever
both	hence	nonetheless	states	whether
buy	her	noone	still	which
but	here	nor	stop	while
by	here's	not	such	whither
c	hereafter	number	sure	who
came	hereby	numbered	t	who'd
can	herein	numbering	take	who'll
can't	hereupon	numbers	taken	who's
cannot	hers	nothing	taking	whoever
caption	herself	now	ten	whole
case	high	nowhere	than	whom
cases	higher	o	that	whomever
certain	highest	of	that'll	whose
certainly	him	off	that's	why
clear	himself	often	that've	will

Table 20. Continued

Stopword List				
clearly	his	old	the	with
co	how	older	their	within
co.	however	oldest	them	without
come	hundred	on	themselves	won't
could	i	once	then	work
couldn't	i'd	one	thence	worked
d	i'll	one's	there	working
did	i'm	only	there'd	works
didn't	i've	onto	there'll	would
differ	ie	open	there're	wouldn't
different	if	opened	there's	x
differently	important	opening	there've	y
do	in	opens	thereafter	year
does	inc.	or	thereby	years
doesn't	indeed	order	therefore	yet
done	instead	ordered	therein	yes
don't	interest	ordering	thereupon	yet
down	interested	orders	these	you
downed	interesting	other	they	you'd
downing	interests	others	they'd	you'll
downs	into	otherwise	they'll	you're
during	is	our	they're	you've
e	isn't	ours	they've	young
each	it	ourselves	thing	younger
early	it's	out	things	youngest
eg	its	over	think	your
eight	itself	overall	thinks	yours
eighty	j	own	thirty	yourself
either	just	p	this	yourselves
else	k	part	those	z
elsewhere	keep	parted	though	
end	keeps	parting	thousand	
ended	kind	parts	thought	

Source: Frakes and Baeza-Yates 1992

LIST OF REFERENCES

Aguilar, F. J. (1967). Scanning the business environment. New York: Macmillan.

Al-Hamad, F. M. (1988). Scanning organizational environments. (Doctoral dissertation, State University of New York at Albany, 1998). Dissertation Abstracts International, 49, 00155.

Apte, C., Damerau, F., & Weiss, S. (1994). Towards language independent automated learning of text categorization models. Proceedings of the 17th Annual ACM/SIGIR Conference on Research and Development in Information Retrieval, pp. 24-30.

Auster, E., & Choo, C. W. (1992). Environmental scanning: Preliminary findings of a survey of CEO information seeking behavior in two Canadian industries. Proceedings of the 55th Annual Meeting of the American Society for Information Science, pp. 48-54.

Auster, E., & Choo, C. W. (1993a). Environmental scanning by CEOs in two Canadian industries. Journal of the American Society for Information Science, 44(4), 194-203.

Auster, E., & Choo, C. W. (1993b). Environmental scanning: Preliminary findings of interviews with CEOs in two Canadian industries. Proceedings of the 56th Annual Meeting of the American Society for Information Science, pp. 246-252.

Badeian, A. G. (1986). Contemporary challenges in the study of organizations. Journal of Management, 12(2), 185-201.

Bajgier, S. M., & Hill, A. V. (1982). An experimental comparison of statistical and linear programming approaches to the discriminant problem. Decision Sciences, 13, 604-618.

Bajwa, D. S., Rai, A., & Brennan, I. (1998). Key antecedents of executive information systems success: A path analytic approach. Decision Support Systems, 22, 31-43.

Bollmann, P., & Wong, S. K. M. (1987). Adaptive linear information retrieval models. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 157-163.

Camillus, J. C., & Datta, D. K. (1991). Managing strategic issues in a turbulent environment. Long Range Planning, 24(2), 67-74.

Chan, W. S. Stock price reaction to news and no-news: Drift and reversal after headlines. Journal of Financial Economics.

Choo, C. W. (1993). Environmental scanning: Acquisition and use of information by chief executive officers in the Canadian telecommunications industry. (Doctoral dissertation, University of Toronto, 1993). Doctoral Abstracts International, 54, 00338.

Choo, C. W. (1995). Information management for the intelligent organization: The art of scanning the environment. Medford, N.J.: Information Today, Inc.

Choo, C. W., & Auster, E. (1993). Scanning the business environment: Acquisition and use of information by managers. In M. E. Williams (Ed.), Annual review of information science and technology, 28. Medford, NJ: Learned Information, Inc.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20, 273-297.

Cronin, M. J. (1993, November/December). What's my motivation? Why businesses are turning to the Internet. Internet World, pp. 40-43.

Crouch, C. J., Crouch, D. B., & Nareddy, K. R. (1990). The automatic generation of extended queries. Proceedings of the Thirteenth International Conference on Research and Development in Information Retrieval, pp. 369-383.

Culnan, M. J. (1983). Environmental scanning: The effects of task complexity and source accessibility on information gathering behavior. Decision Sciences, 14(2), 194-206.

Daft, R. L., Sormunen, J., & Parks, D. (1988). Chief executive scanning, environmental characteristics, and company performance: An empirical study. Strategic Management Journal, 9(2), 123-139.

Daft, R. L., & Weick, K. E. (1984). Toward a model of organizations as interpretations systems. Academy of Management Review, 9(2), 284-295.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. A. (1990). Indexing by latent semantic analyisis. Journal of the American Society for Information Science, 41(6), 391-407.

Denton, D. K. (2001). Better decisions with less information. Industrial Management, 43, 21-25.

Dollinger, M. J. (1984). Environmental boundary spanning and information processing effects on organizational performace. Academy of Management Journal, 27(2), 351-368.

Drucker, P. E. (1998, August 24). The next information revolution. Forbes, pp. 46-53.

Einhorn, H. J. (1980). Overconfidence in judgment. New Directions for Methodology of Social and Behavioral Science, 4(1), 1-16.

Elofson, G., Beranek, P. M., & Thomas, P. (1997). An intelligent agent community approach to knowledge sharing. Decision Support Systems, 20, 83-98.

Fahey, L., & King, W. R. (1977, August). Environmental scanning for corporate planning. Business Horizons, 20(4), 61-71.

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. Journal of Finance, 25(2), 383-417.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188.

Frakes, W. B., & Baeza-Yates, R. (1992). Information retrieval data structures and algorithms. Englewood Cliffs, NJ: Prentice Hall.

Freed, N., & Glover, F. (1981a). A linear programming approach to the discriminant problem. Decision Sciences, 12, 68-74.

Freed, N., & Glover, F. (1981b). Simple but powerful goal programming formulations for the discriminant problem. European Journal of Operational Research, 7, 44-60.

Freed, N., & Glover, F. (1982). Linear programming and statistical discrimination—The LP side. Decision Sciences, 13, 172-175.

Freed, N., & Glover, F. (1986a). Evaluating alternative linear programming models to solve the two-group discriminant problem. Decision Sciences, 17, 151-162.

Freed, N., & Glover, F. (1986b). Resolving certain difficulties and improving the classification power of LP discriminant analysis formulations. Decision Sciences, 17, 589-595.

Gates, A. M. (1990). Environmental scanning in Pennsylvania community colleges: Does it exist? Does it work? (Doctoral dissertation, University of Pittsburgh, 1990). Doctoral Abstracts International, 52, 00239.

Ghoshal, S. (1988). Environmental scanning in Korean firms: Organizational isomorphism in practice. Journal of Business Studies, 19(1), 69-86.

Ghoshal, S., & Kim, S. K. (1986). Building effective intelligence systems for competitive advantage. Sloan Management Review, 28(1), 49-58.

Ghoshal, S., & Whestney, D. E. (1991). Organizing competitor analysis systems. Strategic Management Journal, 12, 17-31.

Glorfield, L. W., & Gaither, N. (1982). On using linear programming in discriminant problems. Decision Sciences, 13, 167-171.

Glover, F. (1990). Improved linear programming models for discriminant analysis. Decision Sciences, 21, 771-785.

Glover, F., Keene, S., & Duea, B. (1988). A new class of models for the discriminant problem. Decision Sciences, 19, 269-280.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). Multivariate data analysis. Upper Saddle River, New Jersey: Prentice Hall, Inc.

Hand, D. J. (1981). Discrimination and classification. New York: John Wiley & Sons Ltd.

Hayes, P.J., & Weinstein, S.P. (1990). Construe/tis: A system for content-based indexing of a database of new stories. Second Annual Conference on Applications of Artificial Intelligence, pp. 1-5.

Henrich, A. (1996). Adapting a spatial access structure for document representations in vector space. Proceedings of the Fifth International Conference on Information and Knowledge Management, pp. 19-26.

Huberman, G., & Regev, T. (2001). Contagious speculation and a cure for cancer: A non-event that made stock prices soar. Journal of Finance, 56, 387-396.

Jain, S. C. (1984). Environmental scanning in US corporations. Long Range Planning, 17(2), 117-128.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Proceedings of the European Conference on Machine Learning, pp. 137-142.

Johnson, L., & Kuehn, R. (1987). The small business owner/manager's search for external information. Journal of Small Business Management, 25(3), 53-60.

Johnson, R. A., & Wichern, D. W. (1982). Applied multivariate statistical analysis. Englewood Cliffs, N. J: Prentice Hall.

Jones, W. P., & Furnas, G. W. (1987). Pictures of relevance: a geometric analysis of similarity measure. Journal of the American Society for Information Science, 38, 420-446.

Jonsen, R. W. (1986) The environmental context for postsecondary education. In P. M. Callan (Ed.), New directions for institutional research: Environmental scanning for strategic leadership, 52. San Francisco: Jossey-Bass.

Keegan, W. J. (1967). Scanning the international business environment: A study of the information acquisition process. (Doctoral dissertation, Harvard University, 1967). Doctoral Abstracts International, X1967, 00001.

Keegan, W. J. (1974). Multinational scanning: A study of the information sources utilized by headquarters executives in multinational companies. Administrative Science Quarterly, 19(3), 411-421.

Klein, H. E., & Linneman, R. E. (1984). Environmental assessment: An international study of corporate practice. Journal of Business Strategy, 5(1), 66-75.

Kleinberg, J., & Tomkins, A. (1999). Applications of linear algebra in information retrieval and hypertext analysis. Proceedings of the Eighteenth ACM SIGOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 185-193.

Kobrin, S. J., Basek, J., Blank, S., & La Palombara, J. (1980). The assessment and evaluation of non-economic environments by American firms. Journal of International Business Studies, 11(1), 32-47.

Koehler, G. J. (1989a). Characterization of unacceptable solutions in LP discriminant analysis. Decision Sciences, 20, 239-257.

Koehler, G. J. (1989b). Unacceptable solutions and the hybrid discriminant model. Decision Sciences, 20, 844-848.

Koehler, G.J. (1991). Improper linear discriminant classifiers. European Journal of Operational Research, 50, 188-198.

Koehler, G. J., & Erenguc, S. S. (1990). Minimizing misclassifications in linear discriminant analysis. Decision Sciences, 2, 63-85.

Koh, C. E., & Watson, H. J. (1998). Data management in executive information systems. Information and Management, 33, 301-312.

Lester, R., & Waters, J. (1989). Environmental scanning and business strategy. London: British Library, Research and Development Department.

Lewis, D. D., & Ringuette, M. (1994). Comparison of two learning algorithms for text categorization. Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval, pp. 81-93.

Lewis, D. D., Schapire, R. E., Callan, J. P., & Papka, R. (1996). Training algorithms for linear text classifiers. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 298-306.

Mangasarian, O. L. (1965). Linear and nonlinear separation of patterns by linear programming. Operations Research, 13, 444-452.

Markowski, C. A., & Markowski, E. P. (1985). Some difficulties and improvements in applying linear programming formulations to the discriminant problem. Decision Sciences, 16, 237-247.

Mayberry, A. (1991). Effects of a selective dissemination of information service on the environmental scanning process of an academic institution. (Doctoral dissertation, University of North Texas, 1991). Doctoral Abstracts International, 52, 00265.

McIntyre, J. R. (1992). A multi-case study of environmental scan as a component of the school planning process. (Doctoral dissertation, Rutgers The State University of New Jersey, 1992). Doctoral Abstracts International, 54, 00291.

Michman, R. D. (1983). Marketing to changing consumer markets: Environmental scanning. New York: Praeger Publishers.

Miller, D., & Friesen, P. H. (1977). Strategy-making in context: Ten empirical archetypes. Journal of Management Studies, 14(3), 253-280.

Moad, J. (1988). The latest challenge for IS is in the executive suite. Datamation, 34(10), 43-52.

Murphy, M. F. (1987). Environmental scanning: A case study in higher education. (Doctoral dissertation, University of Georgia, 1987). Doctoral Abstracts International, 48, 00302.

Newgren, K. E., Rasher, A. A., & LaRoe, M. E. (1984). An empirical investigation of the relationship between environmental assessment and corporate performance. Proceedings of the 44th Annual Meeting of the Academy of Management, pp. 352-356.

Nishi, K., Schoderbek, C., & Schoderbek, P. P. (1982). Scanning the organizational environment: Some empirical results. Human Systems Management, 3(4), 233-245.

O'Connell, J. J., & Zimmerman, J. W. (1979). Scanning the international environment. California Management Review, 22(2), 15-23.

Olsen, M. D., Murthy, B., & Teare, R. (1994). CEO perspectives on scanning the global hotel business environment. International Journal of Contemporary Hospitality Management, 6(4), 3-9.

Pawar, B., & Sharda, R. (1997). Obtaining business intelligence on the Internet. Long Range Planning, 30, 110-121.

Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14 (3), 130-137.

Preble, J. F., Rau, P. A., & Reichel, A. (1988). The environmental scanning practices of US multinationals in the late 1980s. Management International Review, 28(4), 4-14.

Ptaszynski, J. G. (1989). Ed quest as an organizational development activity: Evaluating the benefits of environmental scanning. (Doctoral dissertation, The University of North Carolina at Chapel Hill, 1989). Doctoral Abstracts International, 50, 00440.

Quinlan, J. R. (1993). C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann.

Raghavan, V. V., & Wong, S. K. M. (1986). A critical analysis of vector space model for information retrieval. Journal of the American Society for Information Science, 37(5), 279-287.

Ragsdale, C. T., & Stam, A. (1991). Mathematical programming formulations for the discriminant problem: An old dog does new tricks. Decision Sciences, 22, 296-307.

Raths, D. (1989). The politics of executive information systems. Infoworld, 11(20), 47-52.

Rocchio, J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), The SMART retrieval system: Experiments in automatic document processing (pp. 313-323). Englewood Cliffs, NJ: Prentice-Hall.

Rosen, J. B. (1965). Pattern separation by convex programming. Journal of Mathematical Analysis and Applications, 10, 123-134.

Russell, S., & Prince, M. J. (1992). Environmental scanning for social services. Long Range Planning, 25(5), 106-113.

Salton, G. (1968). Automatic information organization and retrieval. New York: McGraw-Hill Book Company.

Salton, G. (Ed.). (1971). The SMART retrieval system: Experiments in automatic text processing. Englewood Cliffs, NJ: Prentice-Hall.

Salton, G. (1975). A theory of indexing. Regional Conference Series in Applied Mathematics No. 18.

Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of information by computer. Reading, MA: Addison-Wesley Publishing Company.

Salton, G., & McGill, M. J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill Book Company.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

Salton, G., & Yang, C. S. (1973). On the specification of term values in automatic indexing. Journal of Documentation, 29(4), 351-372.

Smeltzer, L. R., Fann, G. L., & Nikolaisen, V. N. (1988). Environmental scanning practices in small businesses. Journal of Small Business Management, 26(3), 55-62.

Subramanian, R., Fernandes, N., & Harper, E. (1993). Environmental scanning in US companies: Their nature and their relationship to performance. Management International Review, 33(3), 271-286.

Subramanian, R., Kumar, K., & Yauger, C. (1994). The scanning of task environments in hospitals: An empirical study. Journal of Applied Business Research, 10(4), 104-115.

Thomas, P. S. (1980). Environmental scanning: The state of the art. Long Range Planning, 13(1), 20-25.

Vapnik, V. N. (1995). The nature of statistical learning theory. New York: Springer.

Walstrom, K. A., & Wilson, R. L. (1997). An examination of executive information systems (EIS) users. Information and Management, 32, 75-83.

Wang, Z. W., Wong, S. K. M., & Yao, Y. Y. (1992). An analysis of vector space models based on computational geometry. Proceedings of the Fifteenth Annual International ACM/ SIGIR Conference on Research and Development in Information Retrieval, pp. 152-160.

Watson, H. J., Rainer, R. K., & Koh, C. E. (1991). Executive information systems: A framework for development and a survey of current practices. MIS Quarterly, 15(1), 13-30.

Wiener, E., Pedersen, J. O., & Weigend, A. S. (1995). A neural network approach to topic spotting. Proceedings of the Fourth Annual Symposium on Document analysis and Information Retrieval, pp. 317-332.

West, J. J. (1988). Strategy, Environmental scanning, and their effect upon firm performance: An exploratory study of the food service industry. (Doctoral dissertation, Virginia Polytechnic Institute and State University, 1988). Doctoral Abstracts International, 50, 00240.

Williamson, D., Williamson, R., & Lesk, M. (1971) The Cornell implementation of the SMART system. In G. Salton (Ed.), The SMART retrieval system: Experiments in automatic text processing (pp. 12-54). Englewood Cliffs, NJ: Prentice-Hall.

Wilson, T. D., & Masser, I. M. (1983). Environmental monitoring and information management in county planning authorities: Representation and exchange of knowledge as a basis of information processes. Proceedings of the 5th International Research Forum in Information Science, pp. 271-284.

Wong, S. K. M., & Yao, Y. Y. (1990). Query formulation in linear retrieval models. Journal of the American Society for Information Science, 41(5), 334-341.

Wong, S. K. M., Yao, Y. Y., & Bollmann, P. (1988). Linear structure in information retrieval. Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 219-232.

Wong, S. K. M., Ziarko, W., Raghavan, V. V., & Wong, P. C. N. (1987). On modeling of information retrieval concepts in vector spaces. ACM Transaction Database System, 12(2), 299-321.

Yang, C. C., Yen, J., & Chen, H. (2000). Intelligent internet searching agent based on hybrid simulated annealing. Decision Support Systems, 28, 269-277.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Information Retrieval, 1(1/2), 69-90.

BIOGRAPHICAL SKETCH

Cheryl Aasheim was born in Miami, Florida. She completed high school in 1987 at Miami Southridge Senior High School. She received her Bachelor of Science in mathematics from the University of Florida in 1991. In 1993, she received her Master of Science in Teaching in mathematics from the University of Florida.

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

Gary J. Koehler, Chair
John B. Higdon Eminent Scholar of
Decision and Information Sciences

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

Sahin Selcuk Erenguc
Professor of Decision and Information
Sciences

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

Haldun Aytug
Assistant Professor of Decision and
Information Sciences

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Doctor of Philosophy.

Jason J. Karceski
Assistant Professor of Finance, Insurance
and Real Estate

This thesis was submitted to the Graduate Faculty of the Department of Decision and Information Sciences in the College of Business Administration and to the Graduate School and was accepted as partial fulfillment of the requirements for the degree of Doctor of Philosophy.

August 2002

_____

Dean, Graduate School